

UNIVERSIDADE FEDERAL DE MINAS GERAIS
Escola de Ciência da Informação
Programa de Pós-Graduação em Gestão & Organização do Conhecimento (PPGGOC)

Francis Bento Marques

EXPLORANDO A PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO NO
BRASIL: um estudo baseado nos dados abertos da CAPES

Belo Horizonte
2025

Francis Bento Marques

**EXPLORANDO A PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO NO
BRASIL: um estudo baseado nos dados abertos da CAPES**

Tese apresentada ao Programa de Pós-Graduação em Gestão & Organização do Conhecimento da Escola de Ciência da Informação da Universidade Federal de Minas Gerais como requisito para obtenção do grau de Doutor em Gestão e Organização do Conhecimento, na área de concentração da Ciência da Informação.

Linha de Pesquisa: Arquitetura & Organização do Conhecimento

Orientadora: Profa. Dra. Benildes Coura
Moreira dos Santos Maculan

Coorientador: Prof. Dr. Renato Rocha Souza

Belo Horizonte
2025

M357e

Marques, Francis Bento.

Explorando a Pós-Graduação em Ciência da Informação no Brasil [recurso eletrônico] : um estudo baseado nos dados abertos da CAPES / Francis Bento Marques . - 2025.

1 recurso online (461 f. : il., color.) : pdf.

Orientadora: Benildes Coura Moreira dos Santos Maculan.

Tese (doutorado)– Universidade Federal de Minas Gerais, Escola de Ciência da Informação.

Referências: f. 425-451.

Apêndice: f. 452-459.

Exigência do sistema: Adobe Acrobat Reader.

1. Ciência da informação – Teses. 2. 1. Pós-graduação – Teses. 3. Bibliometria - Teses. 4. Publicações científicas – Teses. 5. (campo preenchido pela biblioteca) - Teses. I. Maculan, Benildes Coura Moreira dos Santos. II. Universidade Federal de Minas Gerais. Escola de Ciência da Informação. III. Título.

CDU: 002:311

Ficha catalográfica: Maianna Giselle de Paula - CRB: 6/2642

Biblioteca Profª Etelvina Lima, Escola de Ciência da Informação da UFMG



UNIVERSIDADE FEDERAL DE MINAS GERAIS
ESCOLA DE CIÊNCIA DA INFORMAÇÃO - ECI
PROGRAMA DE PÓS-GRADUAÇÃO EM GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO - PPGGOC

FOLHA DE APROVAÇÃO

Explorando a Pós-Graduação em Ciência da Informação no Brasil: Um Estudo Baseado nos Dados Abertos da CAPES

FRANCIS BENTO MARQUES

Tese submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO, como requisito para obtenção do grau de Doutor em GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO, área de concentração CIÊNCIA DA INFORMAÇÃO, linha de pesquisa Arquitetura e Organização do Conhecimento.

Aprovada em 16 de maio de 2025, por videoconferência, pela banca constituída pelos membros:

Profa. Benildes Coura Moreira dos Santos Maculan (Orientadora)
ECI/UFMG

Prof. Renato Rocha Souza (Coorientador)
FGV/RJ

Profa. Mariângela Spotti Lopes Fujita
UNESP

Prof. Flavio Codeço Coelho
EMAp/FGV

Prof. Mauricio Barcellos Almeida
ECI/UFMG

Prof. Roniberto Morato do Amaral
Universidade Federal de São Carlos

Profa. Cristiane Mendes Netto
UNIVALE

Belo Horizonte, 16 de maio de 2025.



Documento assinado eletronicamente por **Benildes Coura Moreira dos Santos Maculan, Professora do Magistério Superior**, em 16/05/2025, às 17:48, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Renato Rocha Souza, Usuário Externo**, em 21/05/2025, às 06:19, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Cristiane Mendes Netto, Usuário Externo**, em 21/05/2025, às 07:04, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Mariângela Spotti Lopes Fujita, Usuário Externo**, em 21/05/2025, às 07:54, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Mauricio Barcellos Almeida, Professor do Magistério Superior**, em 21/05/2025, às 08:34, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Flávio Codeço Coelho, Usuário Externo**, em 21/05/2025, às 12:08, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Roniberto Morato do Amaral, Usuário Externo**, em 21/05/2025, às 20:23, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **4204757** e o código CRC **2011F4EE**.



UNIVERSIDADE FEDERAL DE MINAS GERAIS
ESCOLA DE CIÊNCIA DA INFORMAÇÃO - ECI
PROGRAMA DE PÓS-GRADUAÇÃO EM GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO - PPGGOC

ATA DA DEFESA DE TESE DO ALUNO

FRANCIS BENTO MARQUES

Realizou-se, no dia 16 de maio de 2025, às 14:30 horas, por videoconferência, da Universidade Federal de Minas Gerais, a defesa de tese, intitulada *Explorando a Pós-Graduação em Ciência da Informação no Brasil: Um Estudo Baseado nos Dados Abertos da CAPES*, apresentada por FRANCIS BENTO MARQUES, número de registro 2021651341, graduado no curso de SISTEMAS DE INFORMAÇÃO, como requisito parcial para a obtenção do grau de Doutor em GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO, na área de concentração de Ciência da Informação, à seguinte Comissão Examinadora: Profa. Benildes Coura Moreira dos Santos Maculan - ECI/UFMG (Orientadora), Prof. Renato Rocha Souza - FGV/RJ (Coorientador), Profa. Mariângela Spotti Lopes Fujita - UNESP, Prof. Flavio Codeço Coelho - EMap/FGV, Prof. Mauricio Barcellos Almeida - ECI/UFMG, Prof. Roniberto Morato do Amaral - Universidade Federal de São Carlos, e a Profa. Cristiane Mendes Netto - UNIVALE.

A Comissão considerou a tese:

(X) Aprovada

() Reprovada

Dada a excelência do trabalho apresentado, a Comissão recomendou, por unanimidade, a indicação da tese aos Prêmios Capes de Tese 2025, ANCIB 2025 e UFMG de Teses 2025, reconhecendo sua originalidade, rigor metodológico, relevância científica e contribuição para o avanço da área de Ciência da Informação.

Finalizados os trabalhos, lavrei a presente ata que, lida e aprovada, vai assinada por mim e pelos membros da Comissão.

Belo Horizonte, 16 de maio de 2025.

Assinatura dos membros da banca examinadora:



Documento assinado eletronicamente por **Benildes Coura Moreira dos Santos Maculan, Professora do Magistério Superior**, em 16/05/2025, às 17:48, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Renato Rocha Souza, Usuário Externo**, em 21/05/2025, às 06:19, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Cristiane Mendes Netto, Usuário Externo**, em 21/05/2025, às 07:04, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Mariângela Spotti Lopes Fujita, Usuário Externo**, em 21/05/2025, às 07:54, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Mauricio Barcellos Almeida, Professor do Magistério Superior**, em 21/05/2025, às 08:34, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Flávio Codeço Coelho, Usuário Externo**, em 21/05/2025, às 12:08, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Roniberto Morato do Amaral, Usuário Externo**, em 21/05/2025, às 20:21, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **4204741** e o código CRC **91FBCC8D**.

AGRADECIMENTOS

A DEUS, por ter me permitido chegar até aqui.

Aos meus pais, Francisco e Maria Cristina, por sempre me apoiarem e pela paciência durante o período de estudos.

Ao meu irmão Yuri, por ter me incentivado a iniciar o doutorado e por ter estado sempre junto em todo o percurso.

Às minhas sobrinhas Luna e Yasmin, por terem compreendido o motivo de ficarem tanto tempo sem o Tio.

À minha orientadora, Professora Dra. Benildes Maculan, por confiar em mim, pela compreensão diante das minhas limitações, pela paciência e por estar sempre disposta a me ensinar e ajudar.

Ao Professor Dr. Renato Souza, meu coorientador, pelas orientações e por ter me recebido tão bem em Viena.

Aos meus colegas de trabalho do DTI da UFVJM – Campus Mucuri: Cristiano, Diogo, Eudivane, Jary, Juliano, Matheus, Oséas, Salim e Wallinson, que me apoiaram e cobriram durante o meu afastamento.

Ao Professor Juliano Oliveira, ao Diretor Dr. Jairo Lisboa e à Congregação do ICET/UFVJM, pelo apoio concedido durante o afastamento.

Ao Programa de Pós-Graduação em Gestão e Organização do Conhecimento (PPGGOC) e à Universidade Federal de Minas Gerais (UFMG), representados por seus colaboradores e professores.

Aos colegas do PPGGOC no grupo de WhatsApp, por estarem sempre dispostos a ajudar.

Aos membros das bancas de qualificação e defesa, por aceitarem o convite para participar e contribuírem com apontamentos e sugestões de melhoria que foram fundamentais para a aprovação da pesquisa.

Aos Professores Dr. Gustavo Carvalhal e Dr. Stênio Cavalier, do ICET/UFVJM; ao Professor Dr. João Victor, da FAMMUC/UFVJM; e ao Professor Dr. Flávio Leal, da UFSJ, pelo incentivo a fazer o doutorado e pelo apoio durante o doutoramento.

Aos amigos de outras nacionalidades que conheci em Viena.

À CAPES por ter me concedido a bolsa de doutorado sanduíche por meio do PDSE (Processo nº 88881.980419/2024-01), agradeço também, em nome da minha orientadora, ao

apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq, pelo fomento à pesquisa (Processo nº 307765/2023-7).

Agradeço, ainda, a todas as pessoas que, de alguma forma, contribuíram ou me apoiaram nessa jornada. Meus sinceros agradecimentos!

“Quando você quer alguma coisa, todo o universo conspira para que você realize o seu desejo” (Paulo Coelho - O Alquimista)

RESUMO

A pesquisa teve como tema a produção científica dos programas de pós-graduação *stricto sensu* no Brasil, com foco na área da Ciência da Informação. O estudo, inserido na linha “Arquitetura e Organização do Conhecimento” destaca a importância do uso de dados abertos como instrumento de análise, permitindo observar padrões de crescimento, assimetrias regionais e indicadores que contribuem para a avaliação da ciência nacional. O problema da pesquisa reside na ausência de dados confiáveis e padronizados sobre a área, o que dificulta análises longitudinais e comparativas da produção científica. Para enfrentar essa limitação, o estudo propôs métodos de tratamento e curadoria de dados públicos, buscando construir uma base sólida para análises descritivas e exploratórias. O objetivo geral foi analisar a pós-graduação em Ciência da Informação no Brasil com base em dados abertos de teses e dissertações da CAPES, caracterizando e posicionando o campo em relação ao sistema nacional, por meio de um estudo bibliométrico. Entre os objetivos específicos, destacam-se a identificação dos programas existentes, a observação de sua distribuição regional, o levantamento de dados sobre docentes e discentes e a comparação com o panorama nacional da pós-graduação. A fundamentação teórica abordou os fundamentos da Ciência da Informação, as políticas públicas de pós-graduação, os critérios de avaliação da CAPES, os princípios de dados abertos e as abordagens bibliométricas. A revisão de literatura identificou estudos que apontam o crescimento da pós-graduação no Brasil e na Ciência da Informação, além de desafios como a desigualdade regional e a ausência de padronização dos dados. A metodologia adotada foi exploratória, descritiva e quantitativa, baseada na análise de dados abertos da CAPES. As etapas incluíram coleta, tratamento e normalização dos dados, estruturação em dataframes, recortes temáticos, análise estatística e visualização por gráficos e mapas, com apoio da linguagem Python e bibliotecas como pandas, matplotlib, seaborn e plotly. Foram analisados quatro conjuntos de dados: programas, docentes, discentes e trabalhos defendidos, abrangendo o período de 2004 a 2022. Os dados foram analisados em dois níveis: o panorama nacional da pós-graduação e o recorte específico da Ciência da Informação. Os principais resultados indicam um crescimento expressivo da área de Ciência da Informação em número de programas, na produção científica e no volume de titulados. Contudo, verificaram-se desigualdades regionais, com maior concentração no Sudeste e Sul; variações no tempo médio de titulação; e baixa internacionalização. Foram identificadas falhas nos registros da CAPES, como ausência de identificadores únicos para docentes e discentes, lacunas em campos como resumo e orientação, e inconsistências terminológicas. Esses problemas reforçam a necessidade de padronização e da adoção dos princípios FAIR (Findable, Accessible, Interoperable, Reusable). Como resultado metodológico, foi desenvolvida a MITAB-DA (Metodologia Integrada de Tratamento e Análise Bibliométrica de Dados Abertos), que sistematiza os procedimentos de extração, limpeza, normalização, análise e visualização dos dados da CAPES, oferecendo um modelo replicável para estudos bibliométricos em escala nacional. Ao final, a pesquisa alcança seus objetivos ao oferecer um panorama inédito da pós-graduação em Ciência da Informação, destacando a curadoria de dados como ferramenta estratégica para avaliação institucional e formulação de políticas públicas.

Palavras-chave: Ciência da Informação; Pós-graduação; Dados abertos; Bibliometria; Avaliação científica.

ABSTRACT

This research focuses on the scientific output of *stricto sensu* graduate programs in Brazil, with a specific emphasis on the field of Information Science. The study, inserted in the line “Architecture and Knowledge Organization” highlights the importance of using open data as an analytical tool, enabling the observation of growth patterns, regional asymmetries, and indicators that contribute to the evaluation of national science. The research problem lies in the lack of reliable and standardized data on the field, which hinders longitudinal and comparative analyses of scientific production. To address this limitation, the study proposed methods for processing and curating public data, aiming to build a solid foundation for descriptive and exploratory analyses. The general objective was to analyze graduate education in Information Science in Brazil based on open data from CAPES theses and dissertations, characterizing and positioning the field within the national system through a bibliometric study. Specific objectives included identifying existing programs, examining their regional distribution, collecting data on faculty and students, and comparing Information Science with the broader national graduate education landscape. The theoretical framework covered the foundations of Information Science, public policies for graduate education, CAPES evaluation criteria, open data principles, and bibliometric approaches. The literature review identified studies that point to the growth of graduate education in Brazil and in Information Science, as well as challenges such as regional inequality and lack of data standardization. The methodology was exploratory, descriptive, and quantitative, based on the analysis of open CAPES data. The steps involved data collection, cleaning, and normalization; structuring into dataframes; thematic segmentation; statistical analysis; and visualization through graphs and maps, supported by the Python programming language and libraries such as pandas, matplotlib, seaborn, and plotly. Four datasets were analyzed: programs, faculty, students, and defended works, covering the period from 2004 to 2022. The data were examined on two levels: the national panorama of graduate education and a specific focus on the Information Science field. The main results indicate significant growth in the area of Information Science in terms of the number of programs, scientific production, and number of graduates. However, regional inequalities were observed, with a greater concentration in the Southeast and South; variations in the average time to graduate; and low internationalization. Flaws were identified in CAPES records, such as the absence of unique identifiers for faculty and students, gaps in fields such as abstract and guidance, and terminological inconsistencies. These problems reinforce the need for standardization and the adoption of FAIR (Findable, Accessible, Interoperable, Reusable) principles. As a methodological outcome, the MITAB-DA (Integrated Methodology for the Treatment and Bibliometric Analysis of Open Data) was developed. It systematizes the procedures for extraction, cleaning, normalization, analysis, and visualization of CAPES data, providing a replicable model for bibliometric studies on a national scale. The research is part of the “Knowledge Architecture and Organization” line by focusing on practices for organizing, analyzing, and retrieving scientific information through metric studies. In the end, the research achieves its objectives by offering an unprecedented panorama of graduate studies in Information Science, highlighting data curation as a strategic tool for institutional assessment and public policy formulation.

Keywords: Information Science; Graduate education; Open data; Bibliometrics; Scientific evaluation.

RESUMEN

La investigación tuvo como tema la producción científica de los programas de posgrado *stricto sensu* en Brasil, con énfasis en el área de Ciencia de la Información. El estudio, insertado en la línea “Arquitectura y Organización del Conocimiento”, destaca la importancia del uso de datos abiertos como herramienta de análisis, lo que permite observar patrones de crecimiento, asimetrías regionales e indicadores que contribuyen a la evaluación de la ciencia nacional. El problema de la investigación radica en la ausencia de datos confiables y estandarizados sobre el área, lo que dificulta los análisis longitudinales y comparativos de la producción científica. Para enfrentar esta limitación, el estudio propuso métodos de tratamiento y curaduría de datos públicos, con el objetivo de construir una base sólida para análisis descriptivos y exploratorios. El objetivo general fue analizar el posgrado en Ciencia de la Información en Brasil, a partir de datos abiertos de tesis y disertaciones disponibles en la CAPES, caracterizando y posicionando el campo en relación con el sistema nacional, mediante un estudio bibliométrico. Entre los objetivos específicos se destacan la identificación de los programas existentes, la observación de su distribución regional, el levantamiento de datos sobre docentes y discentes, y la comparación con el panorama nacional del posgrado. El marco teórico abordó los fundamentos de la Ciencia de la Información, las políticas públicas de posgrado, los criterios de evaluación de la CAPES, los principios de datos abiertos y los enfoques bibliométricos. La revisión de la literatura identificó estudios que señalan el crecimiento del posgrado en Brasil y en la Ciencia de la Información, así como desafíos como la desigualdad regional y la falta de estandarización de los datos. La metodología adoptada fue de carácter exploratorio, descriptivo y cuantitativo, basada en el análisis de datos abiertos de la CAPES. Las etapas incluyeron la recopilación, tratamiento y normalización de datos, estructuración en dataframes, recortes temáticos, análisis estadístico y visualización a través de gráficos y mapas, con apoyo del lenguaje de programación Python y bibliotecas como pandas, matplotlib, seaborn y plotly. Se analizaron cuatro conjuntos de datos: programas, docentes, discentes y trabajos defendidos, abarcando el período de 2004 a 2022. Los datos se analizaron en dos niveles: el panorama nacional del posgrado y el recorte específico del área de Ciencia de la Información. Los principales resultados indican un crecimiento significativo en el área de Ciencias de la Información en términos de número de programas, producción científica y número de egresados. Sin embargo, se observaron desigualdades regionales, con mayor concentración en el Sureste y el Sur; variaciones en el tiempo medio de titulación; y baja internacionalización. Se identificaron fallas en los registros del CAPES, como la ausencia de identificadores únicos para profesores y estudiantes, lagunas en campos como resumen y orientación e inconsistencias terminológicas. Estos problemas refuerzan la necesidad de estandarización y la adopción de los principios FAIR (Encontrable, Accesible, Interoperable, Reutilizable). Como resultado metodológico, se desarrolló la MITAB-DA (Metodología Integrada de Tratamiento y Análisis Bibliométrico de Datos Abiertos), que sistematiza los procedimientos de extracción, limpieza, normalización, análisis y visualización de los datos de la CAPES, ofreciendo un modelo replicable para estudios bibliométricos a escala nacional. La investigación se enmarca en la línea “Arquitectura y Organización del Conocimiento” al enfocarse en prácticas de organización, análisis y recuperación de información científica a través de estudios métricos. Al final, la investigación logra sus objetivos al ofrecer un panorama inédito de los estudios de posgrado en Ciencias de la Información, destacando la curación de datos como una herramienta estratégica para la evaluación institucional y la formulación de políticas públicas.

Palabras clave: Ciencia de la Información; Posgrado; Datos abiertos; Bibliometría; Evaluación científica.

LISTA DE FIGURAS

Figura 1 – Rede de colaboração entre países quanto ao tema Inteligência Artificial	89
Figura 2 – Rede de acoplamento bibliográfico a partir de levantamento na Scopus	92
Figura 3 – Delimitação do objeto estudado: universo, objeto e delimitação	123
Figura 4 – Fases da pesquisa bibliométrica	128
Figura 5 – Nome dos campos encontrados nos arquivos do DF1	130
Figura 6 – Nome dos campos encontrados nos arquivos do DF2	132
Figura 7 – Situação dos campos vazios após a execução do script da Quadro 73	216
Figura 8 – Visualização de dados referentes ao nome da IES da última titulação	219

LISTA DE TABELAS

Tabela 1 – Nome dos arquivos que representam o DF1	129
Tabela 2 – Nome dos arquivos que representam o DF2	131
Tabela 3 – Campos existentes no DF1	143
Tabela 4 – Campos existentes no DF2	145
Tabela 5 – Composição dos campos DF3_trabalhos_BR	148
Tabela 6 – Resultado do comando info() da biblioteca Pandas	152
Tabela 7 – Resultado do comando info() da biblioteca Pandas	156
Tabela 8 – Resultado do comando isnull().sum() da biblioteca Pandas	161
Tabela 9 – Resultado da execução da função nunique() da biblioteca Pandas	163
Tabela 10 – Resultado da execução do SCF para o campo ANO_BASE do DF3_trabalhos_BR	165
Tabela 11 – Resultado da execução do SCF para os campos código e nome da grande área no DF3_trabalhos_BR	166
Tabela 12 – Áreas de avaliação encontradas no DataFrame com o uso SCF para o campo NM_AREA_AVALIACAO	168
Tabela 13 – Dados dos trabalhos encontrados por região no DF3_trabalhos_BR	172
Tabela 14 – Dados dos trabalhos encontrados por UF no DF3_trabalhos_BR	173
Tabela 15 – Nomes dos graus acadêmicos encontrados no DF3_trabalhos_BR	177
Tabela 16 – Resultado do script demonstrado no Quadro 43	183
Tabela 17 – Nomes dos graus acadêmicos encontrados no DF3_trabalhos_BR	185
Tabela 18 – Nomes das áreas de avaliação que são diferentes dos códigos	194
Tabela 19 – Total dos dados dos conjuntos de dados referentes aos docentes	196
Tabela 20 – Campos que vão constituir o DF5_docentes_BR	198
Tabela 21 – Dados inseridos no DF5_docentes_BR após a execução do script	200
Tabela 22 – Dados inseridos no DF5 após a execução dos scripts	202
Tabela 23 – Exemplo de dados encontrados após a execução do script que agrupa os IDs aos nomes	207
Tabela 24 – Relação de nomes do último grau de titulação dos docentes	215
Tabela 25 – Dados finais do DF5_docentes_BR após os ajustes realizados	223
Tabela 26 – Total dos dados dos conjuntos de dados referentes aos discentes	225
Tabela 27 – Campos que vão constituir o DF6_discentes_BR	227
Tabela 28 – Resultado do DF6_discentes_BR com os dados de 2004 a 2012	229
Tabela 29 – Resultado do DF6_discentes_BR com os dados de 2004 a 2022	232
Tabela 30 – Dados do campo AN_BASE	233
Tabela 31 – Valores encontrados no campo DS_FAIXA_ETARIA	241
Tabela 32 – Valores encontrados no campo NM_NIVEL_CONCLUSAO_DISCENTE	242
Tabela 33 – Regiões e UF presentes no DF3_trabalhos_BR	248
Tabela 34 – Identificação e o nome do Grau acadêmico presentes no DF3_trabalhos_BR	249
Tabela 35 – Divisões temáticas e conjuntos de dados sobre a avaliação da Pós-Graduação	260
Tabela 36 – Programas de pós-graduação stricto sensu em Ciência da Informação; período 2004-2022	331
Tabela 37 – Características e produção dos programas de pós-graduação em Ciência da Informação no Brasil; dados do período 2004-2022	362
Tabela 38 – Panorama dos programas de Ciência da Informação com início do mestrado e doutorado e produção acadêmica; dados do período 2004-2022	365

LISTA DE GRÁFICOS

Gráfico 1 – Crescimento do número de PPG por grande área: 1998 a 2020	51
Gráfico 2 – Número de Programas de Pós-Graduação – destaque aos anos com ocorrência do Mestrado e/ou Mestrado Profissional	65
Gráfico 3 – Distribuição anual dos estudos selecionados para a revisão	107
Gráfico 4 – Ambientes de coleta de dados	112
Gráfico 5 – Procedimentos de coleta de dados nos estudos da amostra	113
Gráfico 6 – Quantidade de células preenchidas por colunas no DF3_trabalhos_BR após o acoplamento do DF1	154
Gráfico 7 – Quantidade de células preenchidas por colunas no DF3_trabalhos_BR após o acoplamento do DF2	158
Gráfico 8 – Número de trabalhos defendidos na pós-graduação no período de 2004 a 2022	268
Gráfico 9 – Distribuição das Grandes Áreas do Conhecimento em percentual de 2004 a 2022	270
Gráfico 10 – Ranking das Áreas do Conhecimento por ano de 2004 a 2022	271
Gráfico 11 – 20 áreas do conhecimento com maior número de trabalhos de 2004 a 2022	272
Gráfico 12 – 20 IES com o maior número de trabalhos de 2004 a 2022	274
Gráfico 13 – Distribuição dos trabalhos defendidos por regiões do país, no período 2004-2022	275
Gráfico 14 – Distribuição das IES por região, no período 2004-2022	276
Gráfico 15 – Quantidade de trabalhos defendidos por UF, no período 2004-2022	277
Gráfico 16 – Quantidade de trabalhos por grau de titulação, no período de 2004-2022	279
Gráfico 17 – Números de trabalho por grau em cada ano, no período 2004-2022	280
Gráfico 18 – Distribuição percentual dos graus acadêmicos por região, no período de 2004-2022	281
Gráfico 19 – Distribuição das áreas do conhecimento por região, no período 2004-2022	282
Gráfico 20 – Relação entre as grandes áreas do conhecimento e os graus de titulação, período 2004-2022	284
Gráfico 21 – Distribuição das 50 áreas de avaliação por grande área do conhecimento, período 2004-2022	286
Gráfico 22 – Mapa com a nacionalidade dos docentes; período 2004-2022	289
Gráfico 23 – Mapa com a nacionalidade dos docentes sem o Brasil; período 2004-2022	290
Gráfico 24 – Distribuição dos anos de nascimento dos docentes; período 2004-2022	291
Gráfico 25 – Distribuição das categorias dos docentes nas IES; período 2004-2022	293
Gráfico 26 – Distribuição dos títulos dos docentes; período 2004-2022	294
Gráfico 27 – Idade do docente no momento da última titulação; período 2004-2022	295
Gráfico 28 – Distribuição da idade média por titulação dos docentes; período 2004-2022 ...	297
Gráfico 29 – 20 IES com maior número de titulação entre os docentes; período 2004-2022	299
Gráfico 30 – Mapa com países da última titulação do docente; período 2004-2022	300
Gráfico 31 – Mapa com países da última titulação do docente sem o Brasil; período 2004-2022	301
Gráfico 32 – Distribuição das áreas básicas da última titulação dos docentes; período 2004-2022	302
Gráfico 33 – Distribuição das grandes áreas do conhecimento da última titulação dos docentes; período 2004-2022	303
Gráfico 34 – Distribuição dos anos de nascimento dos discentes; período 2004-2022	305
Gráfico 35 – Distribuição da idade do discente no momento da titulação; período 2004-2022	306
Gráfico 36 – Idade do docente no momento da última titulação; período 2004-2022	307

Gráfico 37 – Mapa com a nacionalidade dos discentes; período 2004-2022	308
Gráfico 38 – Mapa com a nacionalidade dos discentes sem o Brasil; período 2004-2022	309
Gráfico 39 – Faixa etária do discente da pós-graduação; período 2004-2022	311
Gráfico 40 – Comparação faixa etária discente doutorado e mestrado apenas acadêmico; período 2004-2022	312
Gráfico 41 – Comparação faixa etária discente doutorado e mestrado, apenas profissional; período 2004-2022	313
Gráfico 42 – Distribuição da quantidade de meses para titulação; período 2004-2022	314
Gráfico 43 – Comparação meses para titulação dos discentes doutorado e mestrado acadêmico; período 2004-2022	316
Gráfico 44 – Comparação meses para titulação dos discentes doutorado e mestrado profissional; período 2004-2022	318
Gráfico 45 – Número de trabalhos da CI defendidos de 2004 a 2022; período 2004-2022 ...	320
Gráfico 46 – Distribuição dos trabalhos defendidos em CI IES; período 2004-2022	321
Gráfico 47 – Distribuição das titulações na CI por regiões do país; período 2004-2022	322
Gráfico 48 – Distribuição das IES por região; período 2004-2022	323
Gráfico 49 – Quantidade de titulações na CI por UF; período 2004-2022	325
Gráfico 50 – Quantidade de trabalhos da CI por grau de titulação; período 2004-2022	326
Gráfico 51 – Números de trabalho da CI por grau em cada ano; período 2004-2022	327
Gráfico 52 – Distribuição percentual dos graus acadêmicos da CI por região; período 2004- 2022	328
Gráfico 53 – Distribuição das IES por grau acadêmico; período 2004-2022	330
Gráfico 54 – Idioma dos trabalhos da CI; período 2004-2022	333
Gráfico 55 – Grau acadêmico dos trabalhos da CI por idioma; período 2004-2022	334
Gráfico 56 – Coocorrência das 20 palavras mais utilizadas nos títulos dos trabalhos da CI; período 2004-2022	335
Gráfico 57 – TOP 20 palavras mais frequentes nos títulos dos trabalhos da CI por ano; período 2004-2022	337
Gráfico 58 – TOP 20 palavras mais frequentes nos trabalhos da CI por grau acadêmico; período 2004-2022	339
Gráfico 59 – TOP 20 bigramas mais frequentes nos títulos dos trabalhos da CI; período 2004- 2022	340
Gráfico 60 – Evolução dos 20 bigramas mais frequentes nos títulos dos trabalhos da CI por ano; período 2004-2022	342
Gráfico 61 – Distribuição dos 20 bigramas mais frequentes nos títulos dos trabalhos da CI por grau; período 2004-2022	343
Gráfico 62 – Coocorrência das 20 palavras mais frequentes nos resumos dos trabalhos da CI; período 2004-2022	345
Gráfico 63 – TOP 20 palavras mais frequentes nos resumos dos trabalhos da CI por ano; período 2004-2022	347
Gráfico 64 – TOP 20 palavras mais frequentes nos resumos dos trabalhos da CI por grau acadêmico; período 2004-2022	349
Gráfico 65 – TOP 20 bigramas mais frequentes nos resumos dos trabalhos da CI; período 2004-2022	350
Gráfico 66 – TOP 20 bigramas mais frequentes nos resumos dos trabalhos da CI por ano; período 2004-2022	351
Gráfico 67 – TOP 20 bigramas mais frequentes nos resumos dos trabalhos da CI por grau acadêmico; período 2004-2022	353
Gráfico 68 – Distribuição dos programas em funcionamento da CI por natureza (status) jurídica; dados do período 2013-2022	355

Gráfico 69 – Distribuição dos programas em funcionamento da CI por dependência administrativa; dados do período 2013-2022	356
Gráfico 70 – Localização das cidades dos programas em funcionamento da CI; dados do período 2013-2022	357
Gráfico 71 – Distribuição dos programas em funcionamento da CI por modalidade; dados do período 2013-2022	358
Gráfico 72 – Distribuição do ano início dos programas em funcionamento da CI; dados do período 2013-2022	359
Gráfico 73 – Distribuição das notas dos programas de Ciência da Informação; dados do período 2013-2022	361
Gráfico 74 – Mapa geral da nacionalidade dos docentes da CI; período 2004-2022	368
Gráfico 75 – Mapa da nacionalidade dos docentes estrangeiros da CI (exceto Brasil); período 2004-2022	369
Gráfico 76 – Distribuição dos docentes da Ciência da Informação por faixas de nascimento; período 2004-2022	370
Gráfico 77 – Distribuição das categorias (Permanente, Colaborador, Visitante) dos docentes da CI nos programas; período 2004-2022	372
Gráfico 78 – Distribuição do último título dos docentes da CI; período 2004-2022	373
Gráfico 79 – Idade do docente da CI no momento da última titulação; período 2004-2022	374
Gráfico 80 – Distribuição da idade média por titulação dos docentes da CI; período 2004-2022	375
Gráfico 81 – As 20 IES com maior número de titulações entre os docentes da CI; período 2004-2022	376
Gráfico 82 – Mapa com países da última titulação do docente da CI; período 2004-2022 ...	377
Gráfico 83 – Mapa dos países estrangeiros da última titulação dos docentes da CI (exceto Brasil); dados do período 2004-2022	378
Gráfico 84 – As 10 áreas básicas da última titulação dos docentes da CI com maior concentração; dados do período 2004-2022	378
Gráfico 85 – Distribuição das grandes áreas do conhecimento da última titulação dos docentes da CI; dados do período 2004-2022	379
Gráfico 86 – Distribuição dos anos de nascimento dos discentes da CI; dados do período 2004-2022	381
Gráfico 87 – Distribuição da idade do discente no momento da titulação da CI; dados do período 2004-2022	382
Gráfico 88 – Idade do docente no momento da última titulação da CI; dados do período 2004-2022	384
Gráfico 89 – Mapa geral com a nacionalidade dos discentes sem o Brasil; dados do período 2004-2022	385
Gráfico 90 – Mapa com a nacionalidade dos discentes da CI sem o Brasil; dados do período 2004-2022	386
Gráfico 91 – Faixa etária do discente da pós-graduação da CI, no momento da titulação, em termos absolutos; dados do período 2004-2022	387
Gráfico 92 – Comparação faixa etária discente doutorado acadêmico, mestrado acadêmico e mestrado profissional; dados do período 2004-2022	388
Gráfico 93 – Distribuição da quantidade do tempo (meses) para titulação da CI; dados do período 2004-2022	389
Gráfico 94 – Comparação meses para titulação dos discentes da CI; dados do período 2004-2022	391

LISTA DE QUADROS

Quadro 1 – Hipóteses projetadas a partir do Plano de Pós-Graduação de 1975	47
Quadro 2 – Quantidade de programas de Pós-Graduação no Brasil por grau em 2020	49
Quadro 3 – Quantidade de programas de Pós-Graduação por grandes áreas: ano 2020	50
Quadro 4 – Quantidade de discentes por grau e por situação, ano base 2020	52
Quadro 5 – Periodização analítica da Pós-Graduação brasileira segundo Almeida (2017)	52
Quadro 6 – Grupos de Trabalho da Ancib e suas ementas	57
Quadro 7 – Programas de Pós-Graduação em Ciência da Informação no Brasil – vigentes em 2022	59
Quadro 8 – Número de cursos oferecidos nos 27 Programas em Ciência da Informação e os anos de implementação do primeiro curso	64
Quadro 9 – Conceitos e aplicações das leis da Bibliometria	79
Quadro 10 – Tipos de redes elaboradas com o uso do VOSviewer	85
Quadro 11 – Tipos de redes de colaboração científica nos estudos bibliométricos	90
Quadro 12 – Resultados iniciais da primeira etapa da revisão integrativa	97
Quadro 13 – Informações específicas das bases de dados consultadas na revisão	98
Quadro 14 – Estabelecimento dos critérios de inclusão e exclusão para seleção dos estudos	100
Quadro 15 – Adequação da string de busca nas bases de dados	101
Quadro 16 – Resultados de estudos pré-selecionados – Análise inicial da terceira etapa	102
Quadro 17 – Aplicação dos critérios de inclusão e exclusão – dados de forma	102
Quadro 18 – Aplicação dos critérios de inclusão/exclusão – dados de conteúdo	104
Quadro 19 – Formação da amostra de análise ao longo das atividades da terceira etapa	104
Quadro 20 – Distribuição das temáticas/objetos dos estudos por áreas e campos do conhecimento, com autores e ano de publicação	109
Quadro 21 – Técnicas de análise de dados e variáveis bibliométricas investigadas	114
Quadro 22 – Resultados por variáveis bibliométricas e proposições para estudos futuros	116
Quadro 23 – Script utilizado para inserir valor null para as células vazias nos arquivos referentes aos anos de 2004 a 2012	136
Quadro 24 – Script para troca de encoding dos arquivos referentes aos anos de 2004 a 2012	137
Quadro 25 – Merge dos arquivos de 2004 a 2012	138
Quadro 26 – Script utilizado para inserir valor null para as células vazias nos arquivos referentes aos anos de 2013 a 2021	140
Quadro 27 – Script para troca de encoding dos arquivos referentes aos anos de 2013 a 2022	141
Quadro 28 – Merge dos arquivos de 2013 a 2022	141
Quadro 29 – Criação do arquivo DF3.csv	149
Quadro 30 – Script para acoplamento do DF1 ao DF3_trabalhos_BR	151
Quadro 31 – Script para acoplamento do DF2 ao DF3_trabalhos_BR	155
Quadro 32 – Script utilizado para entendimento inicial dos dados do DF3_trabalhos_BR	159
Quadro 33 – Script para inserir coluna de índice único no DF3_trabalhos_BR	160
Quadro 34 – Script utilizado para demonstrar os valores únicos de cada coluna do DF3_trabalhos_BR	163
Quadro 35 – Contagem de frequência de cada coluna do DF3_trabalhos_BR	164
Quadro 36 – Script desenvolvido para encontrar divergências entre os códigos e nomes da grande área do conhecimento	167
Quadro 37 – Script desenvolvido para visualizar valores únicos por coluna e linha específicos	174
Quadro 38 – Script desenvolvido para visualizar valores únicos com resultado maior que 1	175

Quadro 39 – Script desenvolvido para visualizar identificadores dos Professores de 2004 a 2012	179
Quadro 40 – Script desenvolvido para visualizar nomes dos orientadores duplicados	180
Quadro 41 – Script desenvolvido para criação do DF4_programas_BR	182
Quadro 42 – Script desenvolvido para alteração do encoding e separador dos arquivos	183
Quadro 43 – Script desenvolvido juntar os dados dos arquivos do conjunto dos programas	183
Quadro 44 – Script desenvolvido para inserção dos códigos nos campos de CD_GRANDE_AREA_CONHECIMENTO e CD_AREA_CONHECIMENTO	186
Quadro 45 – Script desenvolvido para verificação de instituições com nomes diferentes para o mesmo código	187
Quadro 46 – Script desenvolvido para ajustar de instituições com nomes diferentes para o mesmo código	188
Quadro 47 – Script para criar os campos de latitude e longitude ao DF4_programas_BR....	189
Quadro 48 – Script para inserção de valores de latitude e longitude aos campos do DF4_programas_BR	189
Quadro 49 – Script para retornar as IES com mais de um município o mesmo	190
Quadro 50 – Script desenvolvido para inserir o nome do município ao nome da IES	191
Quadro 51 – Script para separar os anos de início dos cursos de mestrado e doutorado	192
Quadro 52 – Script para criar códigos para o nome do grau do programa.	193
Quadro 53 – Script para verificação dos códigos e nomes das áreas de avaliação	193
Quadro 54 – Script desenvolvido para alteração do encoding e separador dos arquivos	199
Quadro 55 – Script desenvolvido para alteração do encoding e separador dos arquivos	200
Quadro 56 – Script desenvolvido para inserção dos dados dos arquivos de 2013 a 2022 no DF5_docentes_BR	201
Quadro 57 – Script desenvolvido para inserção dos dados dos arquivos de 2013 a 2022 no DF5_docentes_BR	203
Quadro 58 – Script desenvolvido para criação de coluna temporária e adição dos nomes dos docentes sem acentos na mesma	204
Quadro 59 – Script desenvolvido para exportação dos nomes dos docentes agrupados pela coluna que não contém acentuação	205
Quadro 60 – Script desenvolvido para remoção de espaços duplos entre strings de um campo determinado previamente	206
Quadro 61 – Script desenvolvido para inserção de ID aos nomes identificados na coluna NM_DOCENTE_S_ACENTO e sem ID	206
Quadro 62 – Script desenvolvido para filtrar nomes diferentes mas com o mesmo ID	207
Quadro 63 – Script desenvolvido alterar os nomes dos docentes pelo ID	208
Quadro 64 – Script desenvolvido para filtrar ID e data de nascimento	209
Quadro 65 – Script desenvolvido para ajustar o ano de nascimento para o ID	210
Quadro 66 – Script desenvolvido para calcular a frequência de cada nacionalidade	212
Quadro 67 – Script desenvolvido para analisar os nomes dos países de nascimento dos docentes	212
Quadro 68 – Script desenvolvido para ajustar os nomes dos países de nascimento dos docentes	213
Quadro 69 – Script desenvolvido para alterar a nacionalidade dos docentes de acordo com o ID	213
Quadro 70 – Script desenvolvido para filtrar as linhas que estão sem valor para o campo DS_TIPO_VINCULO_DOCENTE	214
Quadro 71 – Script desenvolvido para inserir valores faltantes no campo DS_TIPO_VINCULO_DOCENTE	215

Quadro 72 – Script desenvolvido para filtrar as linhas que estão sem valor para o campo NM_ULTIMO_GRAU_TITULACAO	216
Quadro 73 – Script desenvolvido para salvar em um arquivo csv todas as linhas pelo nome do docente sem acento	217
Quadro 74 – Script desenvolvido para inserir valores faltantes no campo DS_TIPO_VINCULO_DOCENTE	218
Quadro 75 – Script desenvolvido para obter os dados do DF5_docentes_BR que contenham AN_ULTIMO_GRAU_TITULACAO menor que 1950 e maior que 2022	218
Quadro 76 – Script desenvolvido para atualizar o campo AN_ULTIMO_GRAU_TITULACAO	219
Quadro 77 – Script desenvolvido para extrair as informações dos campos SG_IES_ULTIMO_GRAU_TITULACAO e NM_IES_GRAU_TITULACAO	219
Quadro 78 – Script desenvolvido para recuperar os nomes das áreas de titulação básica e seus códigos	221
Quadro 79 – Script desenvolvido para inserir dois campos adicionais ao DF5_docentes_BR	222
Quadro 80 – Script desenvolvido para inserir valores aos campos de código e nome de titulação de nível 1	222
Quadro 81 – Script desenvolvido para criação do DF6_discentes_BR	227
Quadro 82 – Script desenvolvido para inserção dos dados de 2004 a 2012	228
Quadro 83 – Script desenvolvido para inserção dos dados de 2013 a 2022	231
Quadro 84 – Script para criação do campo NM_DISCENTE_SEM_ACENTO no DF6_discentes_BR e migração dos valores do NM_DISCENTE	235
Quadro 85 – Script desenvolvido para alterar os valores da coluna NM_PAIS_ORIGEM_DISCENTE	237
Quadro 86 – Script desenvolvido para alterar os valores da coluna NM_PAIS_ORIGEM_DISCENTE	237
Quadro 87 – Script desenvolvido para criar arquivo com o nome dos professores únicos em um csv	238
Quadro 88 – Script desenvolvido para buscar o mesmo valor entre os campos sem acentos no DataFrame DF5_docentes_BR e NM_ORIENTADOR_UNICOS_S_ACENTO.csv	239
Quadro 89 – Script desenvolvido para inserção do campo ID_DOCENTE e alteração do campo NM_ORIENTADOR para NM_DOCENTE	240
Quadro 90 – Script desenvolvido para inserir os valores dos ID_DOCENTE no DF6_discentes_BR	240
Quadro 91 – Script para exportar as linhas vazias de uma determinada coluna	244
Quadro 92 – Script para atualizar linhas no arquivo DF3_trabalhos_BR com base em um csv	245
Quadro 93 – Script desenvolvido para ajustar valores de uma coluna filtrado por duas colunas	246
Quadro 94 – Script desenvolvido para ajustar valores de uma coluna	247
Quadro 95 – Script desenvolvido para filtrar valores por duas colunas	248
Quadro 96 – Script desenvolvido adicionar coluna NM_ORIENTADOR_S_ACENTO	250
Quadro 97 – Script desenvolvido adicionar busca de código do docente entre dataframes	251
Quadro 98 – Script atualizar os nomes dos programas de pós-graduação pelos códigos	252
Quadro 99 – Script desenvolvido para recortar os dados da Ciência da Informação do DF3	254
Quadro 100 – Script desenvolvido para recortar os dados da Ciência da Informação do DF4_programas_CI	255
Quadro 101 – Script desenvolvido para recortar os dados da Ciência da Informação do DF5_docentes_CI5	255

Quadro 102 – Script desenvolvido para recortar os dados da Ciência da Informação do DF6	256
Quadro 103 – Síntese das etapas metodológicas da pesquisa	258
Quadro 104 – Comparação entre esta tese e estudos bibliométricos analisados na revisão de literatura (cap. 2) sobre a produção científica em Ciência da Informação no Brasil	403
Quadro 105 – Síntese da metodologia MITAB-DA	414
Quadro 106 – Síntese dos estudos constantes da amostra para análise	450

LISTA DE ABREVIATURAS

ANCIB - Associação Nacional de Pesquisa e Pós-Graduação em Ciência da Informação
Bvs-Psi Brasil - Biblioteca Virtual em Saúde do Brasil
BDTD - Biblioteca Digital Brasileira de Teses e Dissertações
BRAPCI - Base de Dados Referenciais de Artigos de Periódicos em Ciência da Informação
CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CGEE - Centro de Gestão e Estudos Estratégicos
CNE - Conselho Nacional de Educação
CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico
CTD - Catálogo de Teses e Dissertações
DGA - Dados Governamentais Abertos
EDA - Exploratory data analysis
ENANCIB - Encontro Nacional de Pesquisa e Pós-Graduação em Ciência da Informação
E-LIS - E-Prints in Library and Information Science
FCRB - Fundação Casa de Rui Barbosa
GT - Grupos de Trabalho
IBBD - Instituto Brasileiro de Bibliografia e Documentação
ISTA - Information Science & Technology Abstracts
LGPD - Lei Geral de Proteção dos Dados
LISA - Library and Information Science Abstracts
LISTA - Library, Information Science & Technology Abstracts
LDB - Diretrizes de Bases da Educação Nacional
MBA - Master Business Administration
MEC - Ministério da Educação
NLP - Processamento de Linguagem Natural
ODS - OpenDocument Spreadsheet
OGP - Open Government Partnership
OMS - Organização Mundial da Saúde
PNPG - Plano Nacional de Pós-Graduação
PPG - Programa de Pós-Graduação
PPGCI - Programa Pós-Graduação em Ciência da Informação
PDA - Plano de Dados Abertos

PNE - Plano Nacional de Educação

RACIn - Revista Analisando em Ciência da Informação

SNPG - Sistema Nacional de Pós-Graduação

UEL - Universidade Estadual de Londrina

UF - Unidades da Federação

UFCA - Universidade Federal do Cariri

UFRGS - Universidade Federal do Rio Grande do Sul

UFMG - Universidade Federal de Minas Gerais

UFPA - Universidade Federal do Pará

UnB - Universidade de Brasília

Unesp - Universidade Estadual Paulista

UNIRIO - Universidade Federal do Estado do Rio de Janeiro

WOS - Web of Science

SUMÁRIO

1 INTRODUÇÃO	26
1.1 Motivação, problema e justificativas	29
1.2 Objetivos	34
1.2.1 Objetivo geral	34
1.2.2 Objetivos específicos	34
1.3 Estrutura da tese	35
2 REFERENCIAL TEÓRICO-METODOLÓGICO	37
2.1 A comunicação científica e as comunidades científicas	37
2.1.1 Teses e dissertações como fonte de informação e literatura científica	43
2.2 História da Pós-Graduação no Brasil	45
2.3 A Ciência da Informação no Brasil	54
2.3.1 A Ciência da Informação como ciência social aplicada: programas e cursos de pós-graduação	62
2.4 Estudos bibliométricos na área de Ciência da Informação	68
2.4.1 Fundamentos epistemológicos: bibliometria e curadoria de dados como práticas informacionais	72
2.4.2 As temáticas abordadas	75
2.4.3 Leis, técnicas e tecnologias	78
2.4.4 Visualização de dados	82
2.4.5 Redes de colaboração	87
2.5 Resumo do capítulo	93
3 REVISÃO INTEGRATIVA DA LITERATURA: ESTUDOS BIBLIOMÉTRICOS DA PRODUÇÃO CIENTÍFICA NA CIÊNCIA DA INFORMAÇÃO	95
3.1 Procedimentos da Revisão integrativa da literatura (RIL)	95
3.2 Resultados da RIL	97
3.2.1 Resultado da Primeira etapa da RIL	97
3.2.2 Resultado da Segunda etapa da RIL	99
3.2.3 Resultado da Terceira etapa da RIL	101
3.2.4 Resultado da Quarta etapa da RIL	104
3.2.5 Resultados da Quinta etapa da RIL	106
4 METODOLOGIA	118
4.1 Caracterização da pesquisa	118
4.2 Os dados abertos Capes	119
4.2.1 Dados Abertos Governamentais (DAG)	121
4.3 Delimitação do objeto da pesquisa	122

4.4 Insumos tecnológicos	124
4.5 Etapas e procedimentos metodológicos	127
4.5.1 O estudo bibliométrico	127
4.5.1.1 Etapa 1 - Coleta de dados	129
4.5.1.2 Etapa 2 – Tratamento dos dados	129
4.5.1.3 Etapa 3 - Consolidação dos conjuntos de dados	133
4.5.1.3.1 Fase 1 - Integração dos <i>DataFrames</i>	133
4.5.1.3.2 Fase 2 - Agregação Preliminar dos Arquivos de Dados	135
4.5.1.3.3 Fase 3 - Acoplamento dos arquivos de 2004 a 2012 no DF1	136
4.5.1.3.4 Fase 4 - Acoplamento dos arquivos de 2013 a 2022 no DF2	139
4.5.1.3.5 Fase 5 - Normalização dos dados do DF1 e DF2	143
4.5.1.4 Etapa 4 - Criação do DataFrame (DF3_trabalhos_BR)	147
4.5.1.4.1 Fase 1 - Consolidação dos DataFrames DF1 e DF2 em DF3_trabalhos_BR	150
4.5.1.5 Etapa 5 - Análise exploratória de dados (DF3_trabalhos_BR)	158
4.5.1.5.1 Fase 1 - Entendimento Inicial dos Dados	159
4.5.1.5.2 Fase 2 - Verificação de Valores Ausentes	160
4.5.1.5.3 Fase 3 - Entendimento profundo dos dados	162
4.5.1.5.4 Fase 4 - Preparação dos dados	181
4.5.1.5.5 Fase 5 - Análise exploratória profunda dos dados no DF3_trabalhos_BR	243
4.5.1.6 Etapa 6 – Recorte de dados	253
4.5.1.6.1 Fase 1 - Recorte dos dados da Ciência da Informação no DF3_trabalhos_CI	254
4.5.1.6.2 Fase 2 - Recorte dos dados da Ciência da Informação no DF4_programas_CI	255
4.5.1.6.3 Fase 3 - Recorte dos dados da Ciência da Informação no DF5_docentes_CI	255
4.5.1.6.4 Fase 4 - Recorte dos dados da Ciência da Informação no DF6_discentes_CI	256
4.5.1.7 Etapa 7 – Visualização dos dados	257
4.5.2 Quadro síntese das etapas dos procedimentos metodológicos	258
5 APRESENTAÇÃO DOS RESULTADOS	260
5.1 Resultados da etapa 1 Coleta de dados	260
5.2 Resultados da etapa 2 Tratamento dos dados	262
5.3 Resultados da etapa 3 de Consolidação dos conjuntos de dados	263
5.4 Resultados da etapa 4 de Criação do DataFrame (DF3_trabalhos_BR)	264
5.5 Resultados da etapa 5 Análise exploratória dos dados do DF3	265
5.6 Resultados da etapa 6 – Recorte de dados	266
5.7 Resultado da etapa 7 de Visualização de dados	267
5.7.1 Resultados do conjunto de dados DF3_trabalhos_BR	268
5.7.2 Resultados do conjunto de dados DF4_programas_BR	287
5.7.3 Resultados do conjunto de dados DF5_docentes_BR	288

5.7.4 Resultados do conjunto de dados DF6_discentes_BR	304
5.7.5 Resultados da etapa do Recorte dos dados da Ciência da Informação no DF3_trabalhos_CI	320
5.7.6 Resultados da etapa do Recorte dos dados da Ciência da Informação no DF4_programas_CI	354
5.7.7 Resultados da etapa do Recorte dos dados da Ciência da Informação no DF5_docentes_CI	367
5.7.8 Resultados da etapa do Recorte dos dados da Ciência da Informação no DF6_discentes_CI	381
6 ANÁLISE DOS RESULTADOS	393
6.1 Qualidade dos dados: Discrepâncias e Anomalias	394
6.2 Interpretação dos principais resultados: a Ciência da Informação	396
6.3 Comparação dos resultados com estudos anteriores na CI	400
6.4 Ciência da Informação na Pós-Graduação Brasileira	405
6.5 Resumo do Capítulo	409
6.6 Proposta de metodologia	410
CONCLUSÕES	416
REFERÊNCIAS	424
APÊNDICE A – SÍNTESE DOS ESTUDOS DA QUARTA ETAPA DA RIL	450

INTRODUÇÃO

O crescimento exponencial na geração e disseminação de dados digitais vem alterando profundamente os processos decisórios e as formas de interação social, especialmente nos contextos acadêmico, institucional e governamental (Belluzzo, 2018). A proliferação de dispositivos conectados, como smartphones e sensores, aliada à expansão das redes de comunicação e à popularização das mídias sociais, tem acelerado a produção e circulação de informações em uma escala sem precedentes. Este fenômeno, frequentemente referido como *big data*, tem implicações para a sociedade e representa um desafio e uma oportunidade para a pesquisa científica, a inovação e a tomada de decisões em distintos campos do conhecimento (Deepa *et al.*, 2022).

A capacidade de gerenciar, analisar e extrair valor desses vastos conjuntos de dados tornou-se uma competência valorizada para a competitividade e o avanço do conhecimento (Graves; Hendler, 2014). A análise de grande quantidade de dados envolve a aplicação de métodos estatísticos, algoritmos e técnicas de aprendizado de máquina, tendo em vista identificar padrões e tendências, realizar previsões e gerar insights a partir dos dados. Além disso, questões éticas e legais relacionadas à privacidade, segurança e governança de dados ganham destaque, exigindo abordagens multidisciplinares e a colaboração entre profissionais de diferentes áreas para garantir o uso responsável e eficiente desses recursos informacionais.

Os dados abertos governamentais são um componente essencial no contexto contemporâneo de crescente geração, coleta e disseminação de dados. O movimento crescente de Dados Governamentais Abertos (DGA), do inglês Open Government Partnership (OGP), vem estimulando as organizações governamentais a viabilizarem os seus dados de forma pública, com o intuito de aumentar a transparência (Attard *et al.*, 2015). Caracterizados por serem disponibilizados de forma livre, pública e acessível, esses dados possibilitam que cidadãos, pesquisadores, empresas e organizações utilizem informações produzidas e mantidas pelos governos para diversificados fins. A adoção de políticas de dados abertos visa promover a transparência, a participação cidadã, a inovação e o desenvolvimento socioeconômico, ao mesmo tempo em que contribui para a eficiência e a eficácia das ações governamentais.

No Brasil, o Portal de Dados Abertos do Governo Federal é um exemplo de plataforma que coleta e disponibiliza conjuntos de dados em áreas como saúde, educação, segurança e meio ambiente, permitindo que a sociedade acompanhe e fiscalize o desempenho das políticas públicas e que os pesquisadores acessem informações relevantes para seus

estudos. A Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), que desempenha um papel fundamental no sistema de pós-graduação brasileiro, sendo responsável por fomentar, avaliar e regular os cursos de mestrado e doutorado em programas no país, passou a disponibilizar conjuntos de dados na internet, em consonância com o seu Plano de Dados Abertos (PDA) (Capes, 2017). Os dados abertos da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES têm um papel muito importante, pois possibilitam o acesso a informações detalhadas sobre a qualidade, a produção científica e a internacionalização dos cursos de pós-graduação, bem como dados sobre bolsas de estudo, recursos financeiros e indicadores de desempenho. A disponibilização dessas informações contribui para a tomada de decisões por parte de gestores, pesquisadores, estudantes e outros interessados no avanço da pós-graduação no Brasil.

A transparência e a acessibilidade dos dados abertos da CAPES fomentam o desenvolvimento de estudos e análises que trazem panoramas sob diversas perspectivas que podem levar a melhorias nos programas de pós-graduação e na gestão de recursos educacionais. Além disso, a propagação dessas informações pode permitir o acompanhamento das tendências e do impacto das políticas públicas no ensino superior, favorecendo a identificação de áreas que necessitam de maior atenção e investimento. Dessa forma, os dados abertos da CAPES representam um instrumento valioso para o fortalecimento da qualidade, da inovação e da competitividade da pós-graduação brasileira em âmbito nacional e internacional.

A qualidade dos dados abertos é um aspecto elementar para garantir a confiabilidade, a utilidade e o impacto das informações disponibilizadas por instituições públicas e privadas (Corsar; Edwards, 2017). Dados de baixa qualidade podem levar a conclusões imprecisas, ineficiências na tomada de decisões e, em última instância, a perda de recursos e oportunidades. Portanto, assegurar a qualidade dos dados abertos é uma prioridade para maximizar os benefícios potenciais do acesso e uso dessas informações por cidadãos, pesquisadores, empresas e organizações. A qualidade dos dados abertos engloba aspectos como a acurácia, a completude, a consistência, a atualização e a padronização das informações. A atualização dos dados também é importante para assegurar que as informações reflitam a realidade mais recente, e a padronização se refere ao uso de convenções e padrões comuns que facilitem a interpretação e a integração dos dados (Corsar; Edwards, 2017). Além desses aspectos, a qualidade dos dados abertos também pode ser influenciada por fatores como a documentação, a granularidade e a acessibilidade dos conjuntos de dados. Além disso, é importante promover a colaboração entre desenvolvedores,

pesquisadores e usuários de dados abertos para explorar todo o potencial dessas informações e contribuir para a criação de um futuro mais sustentável e inclusivo (Chen; Chiang; Storey, 2012).

A transformação dos dados abertos por meio de arcabouços como programação, bibliometria, inteligência artificial e processamento de linguagem natural tem potencial para gerar conhecimentos valiosos e facilitar a tomada de decisões em distintos setores da sociedade (Di Vaio; Hassan; Alavoine, 2022). Essas abordagens permitem o desenvolvimento de soluções inovadoras e a geração de insights a partir de grandes volumes de dados, contribuindo para a criação de valor e para a promoção do bem-estar social, econômico e ambiental. De fato, os estudos bibliométricos, que possibilitam quantificar a produção e a comunicação científica, têm contribuído para a identificação das relações estabelecidas no processo de colaboração da ciência e no fortalecimento da comunicação científica como um todo (Melo; Trinca; Maricato, 2021). Assim, segundo esses autores, o uso das técnicas bibliométricas são tendências que visam a avaliar o trabalho de pesquisa que, no Brasil, tem seu enfoque voltado à Pós-Graduação oferecida nas universidades e nos centros de pesquisa do País.

Grácio (2020, p. 20) destaca que os estudos bibliométricos têm por base “conceitos, teorias e procedimentos da Ciência da Informação, da Sociologia da Ciência, da Matemática, da Estatística e da Computação e utilizam procedimentos quantitativos como método de análise”. A natureza prioritariamente quantitativa da bibliometria “tem um papel relevante na análise da produção científica de um país, uma vez que seus indicadores retratam o grau de desenvolvimento de uma área do conhecimento de um campo científico ou do saber” (Araújo; Alvarenga, 2011, p. 51). Desse modo, “a análise da produção científica é importante para todas as áreas do conhecimento, pois propicia um mapeamento do saber construído, o que está sendo feito e o que ainda é preciso fazer” (Schweitzer; Rodrigues, 2013, p. 157). Ao tratar os dados a partir de estudos bibliométricos é necessário um esforço em técnicas de visualização de dados para melhor agregação de valor aos dados compilados e tratados.

A análise visual de dados é uma abordagem poderosa que permite aos pesquisadores e tomadores de decisão explorar, analisar e compreender grandes volumes de informações de maneira eficiente e intuitiva (Brugger *et al.*, 2016). Graves e Hendler (2014), mostram que a visualização de dados pode facilitar a identificação de padrões, tendências e anomalias, bem como a comunicação de resultados complexos para públicos distintos, contribuindo para o avanço do conhecimento e a melhoria das práticas em várias áreas.

No contexto da pós-graduação em geral, a visualização de dados abertos pode oferecer elementos valiosos para melhorar a qualidade, a relevância e a equidade dos programas e atividades de ensino e pesquisa. A análise de dados sobre os pós-graduandos pode ajudar a identificar áreas de força e oportunidade, bem como desafios e barreiras à participação e ao sucesso dos estudantes. Além disso, a visualização de dados pode contribuir para a avaliação e o planejamento de iniciativas e parcerias de pós-graduação, bem como para a definição de prioridades e metas para o desenvolvimento científico e tecnológico. No Brasil, o desenvolvimento da área da Ciência da Informação se concretiza basicamente pelos resultados obtidos em cursos de pós-graduação, razão pela qual o mapeamento e a compreensão do seu processo de produção do conhecimento são, em grande parte, dependentes dos dados de avaliação pela CAPES.

Dentre os diferentes tipos de dados disponibilizados, destacam-se as teses e dissertações como fontes originadas na pós-graduação, por registrarem os resultados das experiências de pesquisa conduzidas pelos discentes. Esses trabalhos, ao serem defendidos, oferecem insumos para a compreensão da produção científica da área, nos âmbitos local, regional, nacional e internacional. A análise dessas fontes pode contribuir para o aprimoramento da Ciência da Informação, orientando o desenvolvimento de políticas, estratégias e ações voltadas à excelência, à inovação e à cooperação entre os programas da área. A seguir, apresentam-se o problema de pesquisa, suas justificativas e os objetivos que estruturam este estudo.

1.1 Motivação, problema e justificativas

A motivação para conduzir esta pesquisa tem origem na trajetória profissional do autor desta tese, graduado em Sistemas de Informação e com mais de duas décadas de atuação na área de informática, em projetos relacionados à programação, banco de dados, análise de dados e levantamento de requisitos. Sempre guiado pelo interesse em transformar dados em informação útil e acessível, identificou, nos dados abertos da CAPES, uma oportunidade concreta de aplicar suas competências técnicas a um contexto científico e socialmente relevante. Optou-se pelo Programa de Pós-Graduação em Gestão & Organização do Conhecimento com o objetivo de integrar competências em computação aos fundamentos teóricos e metodológicos da Ciência da Informação. Assim, a pesquisa reflete o encontro entre uma prática consolidada no campo tecnológico e a busca por novos referenciais para enfrentar os desafios da gestão de dados no âmbito acadêmico.

A quantidade de pesquisas científicas realizadas na área da Ciência da Informação tem crescido nas primeiras décadas do século XXI, cujo efeito desse processo é o aumento da produção do conhecimento e o fortalecimento da área como profissão e disciplina (Kawalec, 2013). Segundo a autora, identificar as tendências de pesquisa, por meio do reconhecimento das temáticas estudadas, é uma estratégia que contribuirá para a compreensão da sociedade da informação e seu desenvolvimento, permitindo que os profissionais e pesquisadores se relacionem mais efetivamente com a área.

Além de identificar as temáticas de investigação, outros aspectos podem ser mensurados e apresentados à comunidade científica e à sociedade, tais como: quantidade de autores, citações, instituições envolvidas, colaboração entre autores, dentre outros (Zupic; Cater, 2015). A caracterização do que é produzido por uma área de conhecimento constitui o campo dos estudos bibliométricos, que visam mensurar a produção científica, a fim de estabelecer os indicadores de qualidade para melhoria do progresso científico (González-Alcaide, 2021).

No Brasil, a produção científica na área da Ciência da Informação está associada, geralmente, ao contexto dos Programas de Pós-Graduação e a literatura relata inúmeras pesquisas (Queiroz; Noronha, 2004), sob os mais diferenciados enfoques, embora pouco se conheça sobre o perfil desses programas e suas produções científicas (Autran *et al.*, 2015). Mapear o que esses programas vêm investigando, como também outros dados que permitam identificar o perfil dos programas é uma tarefa complexa, “[...] uma vez que há múltiplas abordagens, e cujas especificidades apresentam corpo próprio e denso, dificultando essa tarefa [...]”, e esse problema resultará na falta de visibilidade do programa junto à sociedade brasileira e internacional (Araújo; Valentim, 2019, p. 253).

Embora o mapeamento da produção científica dos programas brasileiros de Pós-Graduação por meio dos estudos bibliométricos tenha crescido, muitas abordagens ainda podem ser realizadas, tais como o levantamento da produção por meio de outros canais que não sejam os formais (anais de evento e periódicos) (Moreira; Ferneda, 2020). Outra limitação ainda não superada nos estudos publicados diz respeito à incapacidade de se comprovar “[...] uma relação específica entre volume de pesquisadores e estudantes dos programas de pós-graduação e o volume da produção científica” (Moreira; Ferneda, 2020, p. 142).

Com esse mesmo entendimento, Silva e Oliveira (2021) relatam que as pesquisas têm focado na produção veiculada em periódicos, cujos ambientes de coletas de dados são as bases de dados, sobretudo a Web of Science e a Scopus. Nota-se que a literatura apresenta carência quanto ao desenvolvimento de estudos de abordagem qualitativa e quantitativa que

analise as produções científicas de cada programa de Pós-Graduação, de modo que seja possível acompanhar e comparar a produção do programa com a produção científica da área (Silva, 2021).

É possível perceber que os estudos que mapeiam a produção científica da Ciência da Informação brasileira analisam o que é publicado no formato de artigos (Moreira; Ferneda, 2020; Viana, 2016), cujos trabalhos analisados nem sempre estão relacionados aos programas de Pós-Graduação. Essa constatação revela que ainda há carência de estudar as temáticas investigadas, a partir do produto final defendido junto às bancas avaliadoras, como as teses e dissertações (Viana; Oliveira, 2016). Embora o estudo de Viana e Oliveira (2016) direcionou-se à análise das teses e dissertações, constataram-se problemas não solucionados, como uma análise que abranja todos os programas da área existentes, e que o levantamento seja sem delimitação de tempo.

O problema da delimitação temporal e o limite de análise a uma única instituição já haviam sido identificados e não superados no estudo de Queiroz e Noronha (2004). Os autores acreditam que o perfil completo dos programas de Pós-Graduação e da Ciência da Informação como um todo precisa ser mapeado por meio das investigações bibliométricas, com detalhamento dos temas pesquisados em conjunto com o perfil dos pesquisadores. Em específico, há necessidade de se estudar o perfil dos docentes e de sua produção científica em confronto com os temas das dissertações e teses que orientam, com a finalidade de subsidiar “[...] a avaliação das linhas de pesquisa desenvolvidas nos programas de pós-graduação em Ciência da Informação” (Queiroz; Noronha, 2004, p. 141).

A produção científica tanto dos programas quanto dos pesquisadores (docentes) é um tema em constante desenvolvimento, que se sustenta na apresentação de indicadores como: número de pesquisadores, evolução temporal das publicações (artigos de eventos e de periódicos), índice de produtividade, canais de publicação, colaboração entre pesquisadores, dentre outros (Lança; Amaral; Gracioso, 2018). Mesmo que essa abordagem investigativa não seja inédita, segundo esses autores, o ineditismo se refere ao fato de que a literatura carece de estudos que monitorem o desempenho dos programas. Ou seja, ainda não foram apresentados estudos sobre a evolução das características da produção desses programas no decorrer dos períodos de avaliação da CAPES.

Percebe-se, a partir dos estudos até então publicados na literatura que versam sobre a produção científica dos programas de Pós-Graduação em Ciência da Informação, que o processo de caracterizar a produção científica é uma estratégia de melhoria do desempenho e que enaltece a visibilidade das instituições. No entanto, esse tema ainda não foi explorado em

sua integridade, considerando as várias abordagens de pesquisa que podem ser assumidas e direcionadas a diferentes indicadores de produtividade. Ao analisar, por exemplo, as teses e dissertações defendidas, é possível mensurar um importante indicador ainda pouco explorado, como as relações sociais de colaboração nas bancas de defesa (Telmo, 2019).

A análise do perfil dos docentes dos programas, os temas e a qualidade das produções e as participações colaborativas merecem maior atenção dos estudos, de modo que os interesses de pesquisas sejam explicitados e o processo de pesquisa fortalecido (Temo, 2019). No contexto das participações em bancas, Souza e Maricato (2022) descobriram, ao analisar as bancas formadas na área de Geografia, que a especialidade do membro examinador é um fator decisivo para a composição nessa etapa de avaliação, sendo mais importante que as relações interpessoais. Segundo esses autores, é preciso ampliar as análises, investigando os critérios utilizados para a escolha das bancas de defesas dos programas brasileiros de Pós-Graduação em diferentes áreas do conhecimento, e, nesse processo, identificar as relações entre os docentes e seus interesses de pesquisa.

Como mencionado por Moreira e Ferneda (2020), a importância da produção científica dos programas de Pós-Graduação em Ciência da Informação vem sendo objeto de estudo de muitos pesquisadores. O problema ainda persiste, segundo os autores, no fato de os estudos realizarem levantamentos de natureza meramente quantitativa e, na maioria dos casos, o foco é identificar como a produção é comunicada e em que formato. Essa constatação revela que ainda há lacunas no sentido de estabelecer análises dos indicadores quantitativos e qualitativos, esses últimos detalhando os assuntos do que é pesquisado e, como recomendou Souza e Maricato (2022), levantando o perfil dos pesquisadores, as redes de colaboração e os motivos que levam a formar essas relações.

A análise das temáticas de pesquisa a partir do levantamento das teses e dissertações já foi realizada por Viana (2016), embora não tenha levantado o panorama nacional da Ciência da Informação no Brasil. Ao investigar as temáticas de todos os programas, outros indicadores associados vão sendo desmembrados, como as relações com as linhas de pesquisa e com a formação e interesses dos orientadores e membros de banca. Com essas associações, que vão sendo apontadas, conforme sugerido por Queiroz e Noronha (2004) e Lança, Amaral e Gracioso (2018), possibilita-se a realização de um estudo aprofundado sobre o perfil dos programas de Pós-Graduação em Ciência da Informação do Brasil.

Com o propósito de mapear o perfil dos programas de Pós-Graduação, tendo como sustentação as pesquisas que vão sendo defendidas, mesmo que não publicadas em canais formais de comunicação (anais de evento e periódicos), os principais indicadores que

evidenciem a caracterização dos programas e seus alinhamentos ou correlações com os campos da Ciência da Informação vão sendo identificados. Esses indicadores podem ser, sobremaneira: as linhas de pesquisa, a formação dos docentes, a colaboração nas bancas, os trabalhos de coautoria, as relações com outras áreas do conhecimento (interdisciplinaridade¹), dentre outros. Segundo Moreira e Ferneda (2020) e Silva (2021), essa proposta de investigação ainda não foi realizada e constitui uma estratégia para superar o problema da relação entre produção científica dos programas nacionais e a produção da área da Ciência da Informação.

Assim, conforme recomendado por Queiroz e Noronha (2004), acredita-se que a principal contribuição deste estudo é que será possível estabelecer relações, tanto de aproximação como de distanciamentos entre as temáticas defendidas, as linhas de pesquisa, os interesses dos docentes e a correlação desses construtos com o universo temático da Ciência da Informação (os campos de investigação da área). Por meio dessas correlações, torna-se possível mapear o perfil ou características dos programas e associá-los com a Ciência da Informação, demonstrando as tendências de pesquisa dessa área de conhecimento, a partir dos seus programas de natureza *stricto sensu*.

A partir do exposto, entende-se que apesar da crescente atenção dada à produção científica da área de Ciência da Informação no Brasil, os estudos existentes concentram-se majoritariamente em publicações em periódicos e anais de eventos, desconsiderando, em grande medida, as teses e dissertações como objeto central de análise. Além disso, muitas investigações restringem-se a recortes institucionais ou temporais específicos, o que compromete a obtenção de um panorama abrangente e representativo dos programas de pós-graduação da área. Há, portanto, uma lacuna no mapeamento sistemático da produção científica defendida nos programas *stricto sensu*, especialmente no que se refere à identificação das temáticas abordadas, aos perfis dos docentes e dos discentes da Ciência da Informação. Soma-se a isso a ausência de procedimentos metodológicos consolidados para garantir a confiabilidade, comparabilidade e replicabilidade das análises baseadas nos dados abertos disponibilizados pela CAPES, os quais, embora volumosos, apresentam desafios estruturais e semânticos que demandam tratamento cuidadoso.

A partir desses resultados relatados na literatura, com a identificação do que ainda não foi superado, define-se, para esta pesquisa, o problema a ser investigado, que se orienta pela

¹ Neste estudo não se aprofunda a discussão acerca do caráter interdisciplinar ou não da Ciência da Informação. Porém, respeita-se o aparecimento dessa discussão, sobretudo em citações diretas, visto que a epistemologia da área envolve trazer essas questões, tendo em vista, acredita-se, a própria origem da área, que compreendeu um esforço comum de especialistas de distintas áreas do conhecimento.

seguinte questão de pesquisa: Como mapear e caracterizar a produção científica da pós-graduação em Ciência da Informação no Brasil, com base nos dados abertos da CAPES, e quais procedimentos metodológicos são necessários para garantir a confiabilidade e a replicabilidade dessa análise?

Destaca-se que o problema está embasado nas seguintes situações: crescimento da Ciência da Informação no Brasil, mas pouco levantamento sobre as temáticas das pesquisas defendidas (teses e dissertações), situação essa identificada e não totalmente resolvida no estudo de Viana (2016); melhoria da qualidade dos programas de Pós-Graduação, porém, pouco conhecimento a respeito do número de pesquisas defendidas em cada programa e número de docentes, desafio esse mencionado no estudo de Lança, Amaral e Gracioso (2018); e aumento da colaboração entre docentes, como as participações em banca, no entanto não há estudos que levantem evidências sobre os interesses de pesquisa dos docentes a partir da participação em bancas, sugestão de pesquisa proposta no estudo de Telmo (2019) e Souza e Maricato (2022).

1.2 Objetivos

Este estudo se orienta pelos objetivos descritos a seguir.

1.2.1 Objetivo geral

Analisar a pós-graduação em Ciência da Informação no Brasil com base em dados abertos das teses e dissertações da CAPES, caracterizando o campo e posicionando-o em relação ao sistema nacional por meio de um estudo bibliométrico, buscando estabelecer procedimentos metodológicos de tratamento e organização de dados capazes de assegurar a confiabilidade e a replicabilidade dessa análise

1.2.2 Objetivos específicos

1. Identificar os programas brasileiros de Pós-Graduação na área de Ciência da Informação, tendo em vista verificar a quantidade e a divisão regional.
2. Identificar convergências e divergências entre as regiões brasileiras, para verificar tendências e oportunidades de colaboração.

3. Levantar dados que caracterizem os PPG brasileiros em Ciência da Informação, de modo a permitir análises comparativas entre programas, considerando variáveis como produção acadêmica, distribuição regional e perfil docente.
4. Distinguir os elementos e aspectos que caracterizam a Ciência da Informação brasileira com vistas a permitir relacioná-los aos da pós-graduação no Brasil.
5. Levantar e analisar dados sobre os PPG brasileiros, seus docentes e discentes, para conhecer seus aspectos mais característicos.

Dessa forma, os objetivos delineados neste estudo orientam a análise da pós-graduação em Ciência da Informação no Brasil com base em dados abertos governamentais, com ênfase em teses e dissertações disponibilizadas pela CAPES. A proposta parte do reconhecimento da carência de estudos sistemáticos que mapeiem a produção científica da área com base em sua documentação, identificando padrões regionais, linhas de pesquisa, redes de colaboração e perfis institucionais. Ao empregar um enfoque bibliométrico sobre dados públicos, busca-se não apenas caracterizar os programas e sua produção, mas também oferecer subsídios para o aprimoramento das políticas de avaliação e desenvolvimento da pós-graduação no campo da Ciência da Informação.

1.3 Estrutura da tese

A tese está organizada em sete capítulos, estruturados de modo a articular o percurso teórico, metodológico e analítico da investigação, permitindo compreender a pós-graduação brasileira, com foco na área da Ciência da Informação, a partir dos dados abertos da CAPES. A organização reflete a lógica investigativa adotada, em que cada capítulo desempenha uma função específica na construção dos resultados e no alcance dos objetivos definidos.

Este **Capítulo 1** apresentou o tema, o problema de pesquisa, os objetivos geral e específicos, e a justificativa do estudo. Ao situar a Ciência da Informação como um campo estratégico para a compreensão da produção científica nacional, introduz-se o interesse na relação entre os dados da CAPES e a dinâmica da pós-graduação no país, enfatizando o papel dos dados abertos como insumo para a análise crítica do sistema de formação acadêmica.

O **Capítulo 2** apresenta os fundamentos teóricos e metodológicos da investigação, abordando a caracterização da Ciência da Informação, a pós-graduação e a comunicação científica, assim como articulando os conceitos de organização do conhecimento, ciência aberta, dados abertos e métodos bibliométricos. Também são discutidos os princípios da

avaliação da CAPES e os conceitos relacionados à organização e curadoria de dados, que sustentam as decisões metodológicas do estudo.

O **Capítulo 3** apresenta os trabalhos correlatos, referentes aos estudos bibliométricos realizados sobre a produção na área da Ciência da Informação, identificados por meio de uma revisão integrativa da literatura. A partir dessa revisão, são discutidos estudos nacionais e internacionais que abordam a produção científica na pós-graduação, com destaque para a área da Ciência da Informação. O capítulo contribui para delimitar a lacuna de pesquisa e justificar a originalidade da proposta, ao evidenciar a escassez de análises que combinem curadoria de dados abertos e métodos bibliométricos aplicados às teses e dissertações.

O **Capítulo 4** descreve os procedimentos metodológicos adotados, divididos em etapas articuladas entre si, incluindo coleta dos dados, tratamento e consolidação dos conjuntos de dados e criação de DataFrames estruturados. O capítulo apresenta em detalhe a aplicação prática da curadoria de dados, por meio do uso de Python e bibliotecas como Pandas, Matplotlib e Seaborn, utilizadas para integrar, tratar e visualizar grandes volumes de dados da CAPES.

O **Capítulo 5** apresenta os resultados da pesquisa, organizados conforme as etapas metodológicas. São analisadas as informações referentes aos programas, docentes, discentes e trabalhos defendidos da pós-graduação *stricto sensu* no Brasil entre 2004 e 2022. Também é realizado o recorte da área da Ciência da Informação, com a criação de conjuntos específicos (DF3, DF4, DF5 e DF6) voltados à caracterização da produção científica da área.

O **Capítulo 6** realiza a análise crítica e interpretativa dos dados consolidados, estabelecendo comparações entre a Ciência da Informação e o sistema nacional de pós-graduação. São discutidos padrões, assimetrias, desafios estruturais e aspectos institucionais que impactam o desenvolvimento da área, articulando os achados empíricos com os referenciais teóricos e com as políticas públicas educacionais.

O **Capítulo 7** apresenta as conclusões do estudo, sintetizando os principais resultados, discutindo as contribuições metodológicas e empíricas da pesquisa, apontando suas limitações e sugerindo direções para investigações futuras, especialmente no que se refere à normalização de dados, ao uso de princípios FAIR e à adoção de tecnologias informacionais para a análise científica.

Essa estrutura buscou assegurar progressão lógica, consistência argumentativa e clareza analítica, permitindo que a tese se configure como um estudo representativo da interseção entre Ciência da Informação, estudos métricos e a pós-graduação.

2 REFERENCIAL TEÓRICO-METODOLÓGICO

Neste capítulo apresenta-se o referencial teórico-metodológico, abordando os fundamentos conceituais, teorias e abordagens que delimitam o foco temático desta pesquisa, que é sobre a produção científica dos PPG na área de Ciência da Informação no Brasil. A intenção é expor esses elementos, tendo como base os estudos que investigaram aspectos correlatos à temática ou às informações dispostas em documentos institucionais.

O capítulo explora quatro eixos temáticos que representam o panorama das pesquisas, com o propósito de explicitar o estado da arte sobre o tema e indicar onde ele se situa dentro das áreas de conhecimento estudadas, quais sejam, Ciências Sociais Aplicadas com enfoque na Ciência da Informação. São percorridas, nas seções seguintes, informações sobre a comunicação e as comunidades científicas, a história da Pós-Graduação no Brasil, a Ciência da Informação brasileira e os estudos bibliométricos desenvolvidos por essa área, com enfoque nas temáticas, leis, técnicas, tecnologias, visualização de dados e redes de colaboração.

2.1 A comunicação científica e as comunidades científicas

A comunicação científica corresponde a um tipo específico de comunicação ocorrido entre os indivíduos que realizam as pesquisas científicas (pesquisadores), a partir de propósitos comuns compartilhados por esses indivíduos (Meadows, 1999). Trata-se, segundo citação clássica de Garvey (1979, p. 4, tradução nossa), de um conjunto de atividades relacionadas à produção, à difusão e ao uso de informações sobre uma pesquisa científica, perpassando um processo contínuo, “[...] desde que o cientista tem uma ideia para a sua pesquisa até as informações sobre os resultados desta pesquisa serem aceitos como um componente do conhecimento científico”.

Com esse fundamento, entende-se que a comunicação científica objetiva informar os interesses e resultados das pesquisas desenvolvidas, na expectativa de contribuir para o avanço da ciência, mediante um processo de validação e compartilhamento (Droescher; Silva, 2014). Comunicar a ciência, portanto, é um processo amplo de compartilhamento do conhecimento relacionado às descobertas no campo científico, com a finalidade de estimular a produção de novos conhecimentos e promover impactos nos estudos desenvolvidos pelos pesquisadores (Cooke *et al.*, 2017).

Esse processo é tão antigo quanto a própria história da ciência e representa uma forma de diálogo e interação entre os pesquisadores, tendo em vista a troca de conhecimento que vai sendo aperfeiçoado com os resultados dos estudos (Caribé, 2015). O diálogo e a interação acontecem no decorrer das atividades de produção da pesquisa e os agentes envolvidos, como autores, editores, divulgadores, dentre outros, criam mecanismos para facilitar a troca de informações, de modo que o conhecimento produzido torne-se conhecido, ou seja, no decorrer da comunicação científica o conhecimento é levado a diferentes grupos sociais (Caribé, 2015).

Uma das características principais da comunicação científica é a validação dos resultados obtidos (Campello; Cendón; Kremer, 2000; Silva; Lazzarin, 2019). Mueller (2000) defende que a validação do que é comunicado acontece mediante um processo rigoroso de análise realizado por pares, que possibilita a legitimação do que é descoberto e socializado entre a comunidade de pesquisadores. Assim como Mueller (2000), Silva e Lazzarin (2019) mencionam que a avaliação por pares representa o centro da comunicação científica, mas não é a única atividade do processo, pois é complementada por um conjunto de ações manifestadas em várias etapas, desde a projeção da pesquisa, incluindo o seu desenvolvimento, avaliação, publicação e, por fim, disseminação entre os grupos interessados e a toda a sociedade.

Durante essas etapas, os pesquisadores recorrem aos canais de comunicação, que podem ser classificados como formais e informais (Fernandes; Vilan Filho, 2021). Segundo Meadows (1999), os canais formais apresentam as seguintes características: público potencialmente grande, informação armazenada e recuperável, avaliação prévia, retorno irrisório para o autor, dentre outras. Por outro lado, os canais informais apresentam características opostas e, segundo o citado autor, compreendem, em linhas gerais, os primeiros contatos entre os especialistas para delineamento do que será estudado. Para Meadows (1999), o periódico representa a principal modalidade de canal formal, e, como canais informais, destacam-se as cartas, reuniões científicas, telefonemas, dentre outros.

Fernandes e Vilan Filho (2021, p. 140) relatam que a tramitação das etapas do processo de comunicação científica forma o fluxo da comunicação, definido como “[...] a trajetória das sucessivas comunicações relacionadas a um projeto específico [...]”. Na visão desses autores, o fluxo da comunicação científica constitui o caminho que a pesquisa percorre, desde a sua idealização, o seu registro nos canais informais, a sua publicação oficial em um canal formal (geralmente, um artigo de periódico) até a sua indexação em bases de dados ou outros ambientes que possibilitarão a recuperação do estudo e a sua devida citação em trabalhos futuros.

É possível observar que a comunicação científica, por ser um processo amplo, conforme defendido por Caribé (2015), está sujeita a um conjunto de elementos ou fatores que se inter-relacionam, tais como: autores, instituições, canais de comunicação, etapas, fluxos, dentre outros. Em conjunto, esses elementos formam o que a literatura denomina de sistema de comunicação científica, ou seja, uma rede de relacionamentos estabelecida entre diferentes agentes e com uso de ferramentas tecnológicas que facilitam a transmissão do conhecimento produzido pela ciência (Campello; Cendón; Kremer, 2000).

Assim, os sistemas de comunicação científica são modelos utilizados para garantir o gerenciamento da produção científica, incluindo a sua disseminação para uso na sociedade, com ferramentas, processos de trabalho e metodologias compartilhadas por grupos de pesquisadores (Coelho; Brandão, 2019). Na visão desses autores, os sistemas passam por transformação, ao longo do tempo, embora duas características básicas permanecem, a saber: 1 – os cientistas são responsáveis pela geração de conteúdo (como autores); e 2 – é garantido o controle da qualidade (por meio de editores e revisores). Neste sentido,

O sistema de comunicação científica, que é subjacente à comunidade científica, é constituído por diversos elementos e processos, entre os quais seus atores e respectivas funções e atividades. É o resultado das funções e atividades de pesquisadores, editoras, bibliotecas, agências de fomento e universidades que contribuem para que a produção, compartilhamento e uso do conhecimento de modo que a ciência possa ser viabilizada (Freitas; Leite, 2019, p. 273).

De fato, ao comunicar os resultados da pesquisa, por meio dos elementos do sistema de comunicação científica, os grupos de pesquisadores compartilham suas ideias e incorporam novos artefatos e práticas, de modo que o fluxo da comunicação seja constantemente adaptado às necessidades atuais. Isso evidencia a capacidade adaptativa do sistema de comunicação, desde que as mudanças sejam corroboradas pelas comunidades científicas ou grupos de pesquisadores (Cooke *et al.*, 2017).

As comunidades científicas representam um elemento essencial do processo de comunicação científica e de estruturação do sistema de comunicação, pois são elas que validam ou aprovam o conhecimento produzido (Campello; Cendón; Kremer, 2000). Para Guimarães *et al.* (2018, p. 523), as comunidades científicas são grupos que se associam e compartilham, além dos resultados de pesquisa, “[...] valores, normas de conduta e comportamentos comuns, isto é, padrões institucionalizados sobre como produzir conhecimentos e tecnologias em determinado campo do saber [...]”. Esses grupos se fortalecem, tornam-se autorregulados, legítimos e são valorizados, à medida que “[...] delimitam espaços de atuação e estabelecem como fazer ciência no campo de interesse, tendo

por base determinados padrões de conduta transversais, aplicados a distintos campos de conhecimento científico [...]” (Guimarães *et al.*, 2018, p. 523).

A característica principal das comunidades científicas é o reconhecimento dos interesses de pesquisa entre seus membros e a iniciativa adotada na contribuição recíproca para a troca e produção de conhecimentos. No que tange ao modo de produzir conhecimento, os integrantes compartilham os mesmos fundamentos, ou seja, adotam regras metodológicas que obrigam ao cientista fornecer evidências empíricas e consistência lógica nos resultados, cujos procedimentos estarão sustentados “[...] por normas morais apoiadas socialmente e possibilitam à ciência cumprir sua meta institucional de ampliação do conhecimento certificado [...]” (Mattedi; Spiess, 2010, p. 77).

Conforme Choo (1998), as organizações — incluindo as acadêmicas — atuam como sistemas de construção de sentido (*sensemaking*), criação de conhecimento e deliberação, em um ciclo contínuo de interpretação e ação. A informação, nesse contexto, não é apenas transmitida, mas interpretada, negociada e formalizada, o que permite à comunidade científica adaptar-se às mudanças em seu ambiente e manter-se coesa em torno de valores e objetivos comuns.

A formação de uma comunidade científica é um processo amplo, moroso e detalhado, cuja principal motivação, na maioria das vezes, está relacionada à identificação de semelhança entre os perfis dos pesquisadores, sobretudo no que tange às temáticas estudadas e aos interesses de cada pesquisador (Dias; Moita; Dias, 2019). No contexto brasileiro, a formação das comunidades científicas está relacionada, sobremaneira, à instituição universitária, cujas atividades foram ampliadas a partir de 1990, e suas atividades também acompanharam a dilatação, “[...] seja por meio do polo do ensino com a criação de novos cursos e de novas vagas, seja através do polo da pesquisa com o fortalecimento e a constituição de novos grupos de pesquisa, seja ainda pelo pilar da extensão e dos projetos que o sustentam” (Silva; Lopes, 2020, p. 5).

Assim, de acordo com Furtado (2015), o investimento no ensino superior e o fortalecimento das atividades de pesquisa oportunizaram a especialização dos pesquisadores, cujo resultado desse processo foi o estímulo do trabalho em equipe, a aproximação entre os membros das comunidades, considerando o perfil de cada um e o interesse por temas específicos. Nesse sentido, as comunidades científicas passam a se caracterizar pela diversidade de seus membros, os quais possuem formações distintas, mas pontos de semelhança nos interesses de pesquisa, o que favorece o desenvolvimento de novas temáticas

e de práticas interdisciplinares, além de enaltecer o trabalho colaborativo (Gonçalves; Karpinski; Araújo, 2021).

No entendimento de Anjos e Rodrigues (2019), as comunidades científicas brasileiras apresentam-se com perfis de pesquisadores altamente diferenciados, visto que a atividade científica, no País, é relativamente nova, em conjunto com as práticas interdisciplinares e com o processo de colaboração científica. Embora as formações desses membros também sejam diferentes, nota-se que há convergência colaborativa entre diferentes membros, em especial, no que tange ao estabelecimento de grupos de pesquisa voltados para temas específicos e à tendência à colaboração, com participação de membros externos ao Brasil (Santana, 2015). Esse estudo também descobriu que a colaboração externa é um comportamento inerente aos grupos de pesquisa brasileiros e que os pesquisadores mais produtivos são os atores que mais agregam membros exteriores.

Souza *et al.* (2022) corroboram parte dos resultados de Santana (2015) ao destacarem que a produção científica das áreas de conhecimento concentra pesquisadores de áreas distintas e que a prática da colaboração tem crescido em todas as comunidades científicas brasileiras. A região Sudeste do País, conforme dados do citado estudo, é referência quanto ao número de publicações, com destaque, também, para o número de cursos de formação, prevalecendo o Mestrado em detrimento ao Doutorado, cujos pesquisadores atuam na realização de múltiplas atividades integradas, a saber: busca por financiamento, iniciação científica, publicações, orientações, ensino, elaboração de projetos de pesquisa, atualização de plataformas, prestação de contas, entre outras.

Dentre as atividades desenvolvidas pelos pesquisadores em suas comunidades científicas, destaca-se a colaboração no processo de pesquisa, que se manifesta quando há interesses semelhantes de pesquisa entre os membros (Dias; Moita; Dias, 2019). Para esses autores, no processo de colaboração, as relações são estabelecidas a ponto de formarem as redes de colaboração, observando-se que as grandes áreas do conhecimento desenvolvem seus trabalhos de modo mais colaborativo em comparação às áreas especializadas. O referido estudo também constatou que, no caso dos pequenos grupos de pesquisa, eles têm trabalhado em colaboração, o que tende a aumentar o número de publicações de seus pesquisadores, visando à participação conjunta dos mesmos coautores, geralmente das mesmas áreas de atuação e das mesmas instituições.

As redes de colaboração podem ser definidas como as interações estabelecidas entre diferentes membros de uma ou mais comunidades científicas, mediante a prática de um trabalho conjunto, tendo em vista interesses ou objetivos comuns entre esses agentes

(Carneiro; Ferreira Neto; Santos, 2020). O efeito desse processo, conforme esses autores, é a atuação conjunta, considerada um elemento fundamental para o desenvolvimento de projetos que ampliem a produtividade científica, com o propósito de obter avanços recíprocos para os distintos campos do conhecimento. Ademais, ao trabalhar em redes, há possibilidade de “[...] potencializar o processo de internacionalização da ciência por meio da inserção dos pesquisadores em redes de colaboração mundiais” (Carneiro; Ferreira Neto; Santos, 2020, p. 147).

A formação das redes é uma ocorrência constante, em especial com o avanço tecnológico, que proporcionou novos canais de interação, encurtando a distância entre os pesquisadores. Em linhas gerais, as comunidades científicas e seus pesquisadores optam por trabalhar em redes, na busca pelo aumento da visibilidade e pelo reconhecimento pessoal (Santos; Pereira, 2019). Além dessas vantagens, Santos e Pereira (2019) elenca outras seis, a saber: 1 - maior produtividade, racionalização do tempo e uso da mão de obra científica; 2 - probabilidade maior de conseguir financiamento e recursos para pesquisa; 3 - profissionalização da ciência; 4 - realização de pesquisas multidisciplinares; 5 - treinamento para pesquisadores e orientandos; e 6 - maior divulgação das pesquisas.

A colaboração tem sido praticada em todas as áreas do conhecimento e compreende um “[...] trabalho conjunto entre os diferentes atores, por exemplo: pesquisadores, países, universidades, departamentos, temáticas, linhas de pesquisa [...]” (Alves; Dalessandro; Santos, 2019, p. 139). Justifica-se a formação das redes de colaboração pelas comunidades, pois elas desempenham um papel fundamental que está relacionado à difusão, troca e transmissão do conhecimento e informação entre os diferentes participantes da comunicação científica. Logo, entende-se que a colaboração é uma prática que influencia a produtividade científica e as políticas que envolvem prestígio e visibilidade científica, tanto dos pesquisadores, das instituições e/ou das comunidades científicas (Alves; Dalessandro; Santos, 2019).

Na área da Saúde, por exemplo, em específico no campo da Medicina Veterinária, a colaboração tem se intensificado, cujos pesquisadores desenvolvem trabalhos em conjunto (coautorias) ou realizam parcerias em projetos de pesquisa instituídos no âmbito interno ou externo à instituição em que atuam (Silva; Targino, 2018). O resultado dessa colaboração é a formação de redes, com trocas de conhecimentos e maior visibilidade e prestígio dos pesquisadores ao terem seus estudos citados, embora os mais produtivos não sejam, necessariamente, os mais citados (Silva; Targino, 2018).

Na área da Comunicação, as redes de colaboração também são identificadas, com propósitos parecidos às demais áreas, desde as justificativas de cunho pessoal do pesquisador

(como o aumento da produtividade) até os aspectos mais específicos, envolvendo o progresso e o desenvolvimento da ciência (como necessidade de opiniões externas para avançar as descobertas) (Cajazeira; Silva, 2021). A fim de analisar profundamente as colaborações, os autores citados recorreram às técnicas bibliométricas, identificando a diversidade de instituições envolvidas, pesquisadores de diferentes campos e autoria compartilhada.

2.1.1 Teses e dissertações como fonte de informação e literatura científica

Mueller (2000, p. 22) destaca que “a produção da literatura de uma área científica envolve muitas e diferentes atividades de comunicação entre os pesquisadores, algumas das quais antecedem e outras se seguem a sua publicação”. Em sua origem a literatura científica se resumia a cartas, livros e alguns escritos individualmente publicados, com o objetivo de compartilhar suas pesquisas. As tábuas de argila, ainda no século VI a.C., podem ser consideradas literatura científica, uma vez que registravam, por exemplo, as descobertas referentes a eclipses lunares e a importantes eventos astronômicos. Os resultados dos estudos de Galileu Galilei, por sua vez, foram registrados à mão em um conjunto de cartas, e, ademais, os princípios da matemática postulados por Isaac Newton foram publicados em uma série de livros em 1686.

Após meados do século XVI a literatura científica, em sua grande maioria, passou a ser publicada em artigos de periódicos revisados por pares, ainda que os livros e cartas ainda sejam utilizadas para o registro do conhecimento científico. Os periódicos científicos se ampliaram no mundo inteiro e, em meados do século XVIII, a maioria das grandes cidades europeias tinha sua própria sociedade científica, cada uma com sua própria publicação científica.

Com a expansão de publicação de literatura científica ficou mais difícil acompanhar as descobertas mais recentes, razão pela qual no século XVIII alguns periódicos incluíam nas suas edições os resumos de trabalhos de pesquisa científica publicados em outros periódicos. Já em 1945, Vannevar Bush vislumbrou um repositório para armazenar toda a literatura científica mundial, de modo que pudesse ser acessado por qualquer indivíduo, pois, para o autor, a literatura científica só é útil se for consultada.

Le Coadic (1996, p. 27), que discute sobre o objeto e o conteúdo dos estudos em Ciência da Informação, afirma que:

As atividades científicas e técnicas são o manancial de onde surgem os conhecimentos científicos e técnicos que se transformarão, depois de registrados, em informações científicas e técnicas. Mas, de modo inverso, essas atividades só

existem, só se concretizam, mediante essas informações. A informação é o sangue da ciência. Sem informação, a ciência não pode se desenvolver e viver. Sem informação a pesquisa seria inútil e não existiria o conhecimento.

Discorrendo sobre o ambiente de bibliotecas, Grogan (2001) declara que a informação científica está centrada em cinco fontes básicas: o catálogo, as fontes bibliográficas, as obras de consulta, a literatura especializada e a própria equipe de bibliotecários. O autor destaca que a literatura especializada é a mais relevante e “compreende a literatura primária, como as monografias originais, periódicos especializados, anais de conferências, relatórios de pesquisas, publicações oficiais, teses² [e dissertações], manuscritos e material de arquivo”, assim como “a literatura secundária na forma dos livros e periódicos mais conhecidos” (Grogan, 2001, p. 117).

Uma maior visibilidade para as teses e dissertações ocorreu no Brasil com a iniciativa da CAPES com a publicação da Portaria 13, de 2006, que estabeleceu a obrigatoriedade da divulgação das teses e dissertações defendidas em PPG *stricto sensu* no país. Para atender à Portaria foi necessário a implantação de bibliotecas digitais de teses e dissertações (BDTD) pelas instituições, para disponibilizar a produção acadêmica em formato digital, com vistas a auxiliar a recuperação do conhecimento acumulado nesses documentos.

Essa iniciativa da CAPES encontra respaldo em Carelli (2004, on-line), quando a autora afirma que a ciência é uma “atividade institucional” e, como tal, é “produzida normalmente, nos institutos de pesquisa e nas universidades, principalmente nos cursos de pós-graduação evidenciando estreito vínculo entre atividades de pesquisa e programas de pós-graduação”. A autora ainda destaca que “torna-se oportuno realizar estudos de produção científica” (Carelli, 2004, on-line) no contexto da pós-graduação.

Isso encontra respaldo na visão de Choo (1998), quando afirma que os documentos científicos, como teses e dissertações, são mais do que registros formais, pois eles representam materializações do processo interpretativo e colaborativo das comunidades epistêmicas. Ao documentar e compartilhar os resultados das investigações, os cientistas constroem interpretações compartilhadas, convertem conhecimento tácito em conhecimento explícito e fornecem insumos para ações futuras. Assim, a comunicação científica constitui a infraestrutura simbólica e informacional que sustenta a continuidade e a renovação do conhecimento dentro das comunidades científicas.

² Destaca-se que Denis Joseph Grogan nasceu na Inglaterra e lá “tese (thesis) é normalmente utilizado para descrever todo o gênero, independentemente do grau acadêmico a que se refere” Cendón, 2007, on-line).

2.2 História da Pós-Graduação no Brasil

A Pós-Graduação é um nível educacional que se propõe a aperfeiçoar ou aprofundar a formação profissional adquirida nos Cursos de Graduação, correspondendo, portanto, ao nível mais elevado do ensino superior brasileiro (Almeida, 2017). Tal nível educacional abrange os programas de Mestrado e Doutorado, os cursos de Especialização, de Aperfeiçoamento e outros, abertos a candidatos diplomados em cursos de Graduação e que atendam às exigências das instituições de ensino (Brasil, 1996).

Embora a Pós-Graduação tenha sido oficializada na Lei de Diretrizes de Bases da Educação Nacional (LDB), no ano de 1996, Pilati (2006) afirma que o marco histórico inicial da Pós-Graduação como modalidade de ensino está no Decreto de n. 29.741, de 11 de julho de 1951, legislação essa que representa o berço de nascimento da CAPES. Os objetivos de criação da Capes naquela época foram: 1 - assegurar a existência de pessoal especializado em quantidade e qualidade suficientes para atender às necessidades dos empreendimentos públicos e privados que visam ao desenvolvimento econômico e social do País; e 2 - oferecer aos indivíduos mais capazes, sem recursos próprios, acesso a todas as oportunidades de aperfeiçoamento (Brasil, 1951).

Nota-se que a preocupação principal da Capes está na formação de recursos humanos para atuar na educação superior, cujo Decreto mencionado dispôs sobre os parâmetros para a oferta dos primeiros cursos de Pós-Graduação do Brasil, com enfoque na formação continuada e de natureza *lato sensu*³. As disposições para esses cursos foram ampliadas no ano de 2007, a partir da Resolução n. 1, de 8 de junho de 2007, do Ministério da Educação (MEC), ao estabelecer que tais cursos visam ao aperfeiçoamento profissional e são destinados a candidatos diplomados em cursos de Graduação ou demais cursos superiores (Brasil, 2007).

Ainda no decorrer da década de 1950, segundo Pilati (2006), foi instituído pela Capes o Programa Universitário, com o objetivo de acelerar a contratação de professores universitários, tendo em vista a carência desses profissionais no País. Segundo o autor, o Programa tinha a proposta de firmar contratos com professores visitantes, a maioria estrangeiros, para realização de atividades de intercâmbio e cooperação, com a oferta de

³ A expressão latina *lato sensu* significa, literalmente, "em sentido amplo", e hoje compreendem programas de especialização e cursos tais como Master Business Administration (MBA), sendo que, se aprovado, o aluno obterá certificado. Por sua vez, a expressão latina *stricto sensu* significa, literalmente, "em sentido específico", e hoje compreendem programas de Mestrado e Doutorado, sendo que, se aprovado, o aluno obterá diploma (Brasil, 2018, on-line).

bolsas de estudos, concessão de eventos científicos, abrangendo todas as áreas de conhecimento até então instituídas no Brasil.

Salienta-se que até meados de 1960, a Pós-Graduação esteve restrita aos cursos de natureza *lato sensu*, sendo que uma nova legislação foi implementada, promovendo mudanças profundas no ensino da Pós-Graduação. Assim, por meio do Parecer n. 977, de 3 de dezembro de 1965, o MEC determinou a existência dos cursos de natureza *stricto sensu* (Pilati, 2006). A redação desse parecer trouxe os fundamentos da Pós-Graduação do modo como é concebida na atualidade (segunda década de XXI), sustentada por dois grandes tipos de curso: Mestrado e Doutorado. Diferente dos cursos de Especialização, o Mestrado e o Doutorado exigem a realização de pesquisas avançadas e complementares a temas estudados na Graduação (Almeida, 2007).

Esse parecer foi alterado por meio da Resolução n. 24, de 18 de dezembro de 2002, do Conselho Nacional de Educação (CNE), ao mencionar que as instituições de ensino superior são autônomas para a criação de cursos de Pós-Graduação, mas devem formalizar os pedidos de reconhecimento dos novos cursos por elas criados até, no máximo, 60 dias após ato formal de criação por seus conselhos superiores. Ademais, o texto da resolução exigiu que os cursos de Mestrado e Doutorado fossem instituídos, apenas, mediante autorização do MEC (Brasil, 2002).

Como se nota, no período de 1950 a 1965, o Brasil regulamentou a legislação para a Pós-Graduação, embora poucos resultados foram alcançados, sobretudo por causa da falta de investimentos e de políticas públicas (Fernandes, 1975). Essa situação começa a melhorar, a partir das políticas educacionais para ampliação do acesso à educação, na década de 1970 (Morosini, 2009). De acordo com a autora, o objetivo dessas políticas foi democratizar o acesso ao ensino e à pesquisa, por conseguinte, eliminar as desigualdades geradas com o período ditatorial ocorrido nas décadas anteriores. A pesquisadora acrescenta que, como modalidade educacional, a Pós-Graduação foi planejada para formar e qualificar os recursos humanos atuantes em ciência e tecnologia, com o objetivo de garantir o desenvolvimento da nação (Morosini, 2009).

Segundo Nobre e Freitas (2017), nas últimas décadas do século XX, iniciativas governamentais foram intensificadas em favor da educação brasileira, com os movimentos de reestruturação das universidades e ampliação de cursos superiores, reservas de vagas, dentre outras conquistas. Em específico, para a Pós-Graduação, um marco histórico está no ano de 1975, durante o governo do militar Ernesto Beckmann Geisel, com a implementação do I Plano Nacional de Pós-Graduação (PNPG), seguido pela reformulação e envolvimento de

diferentes instituições, como: o MEC, a CAPES e o CNPq, instituições essas que se consolidaram como gerenciadoras dos recursos destinados à Pós-Graduação, por conseguinte, avançando o processo de pesquisa no Brasil (Nobre; Freitas, 2017).

O Plano Nacional de Pós-Graduação (PNPG) é a política de governo que direciona as atividades do conjunto de instituições de ensino superior que compõem o Sistema Nacional de Pós-Graduação (SNPG), como as instituições públicas e privadas, confessionais e comunitárias. O PNPG integra o Plano Nacional de Educação (PNE) do MEC que determina diretrizes, metas e estratégias para a política educacional (Barbosa, 2020, p. 1).

O I PNPG foi implementado, a princípio, no ano de 1975, dispondo um conjunto de diretrizes para regulamentação e funcionamento dos cursos de Pós-Graduação brasileiros, tais como: diretrizes institucionais, estabilização financeira, padrões de desempenho, processos de seleção, credenciamento, planejamento, dentre outras. Destaca-se que o Plano se manifestou como uma política pública, com a pretensão de evoluir para uma nova etapa do sistema universitário, durante a qual as atividades de Pós-Graduação assumiram importância estratégica crescente para o desenvolvimento científico, econômico e tecnológico do País (Hostins, 2006).

Pela análise ao Plano de 1975, nota-se o seu teor ambicioso, no sentido de avançar a Pós-Graduação como uma nova e importante modalidade educacional. Assim, foram elencadas três hipóteses ou metas a serem alcançadas, nos anos seguintes, conforme apresentado no Quadro 1.

Quadro 1 – Hipóteses projetadas a partir do Plano de Pós-Graduação de 1975

Hipótese ou diretrizes	Síntese descritiva
1 – Integração ensino e pesquisa	O ensino e a pesquisa devem estar integrados em todos os níveis e os vários níveis devem estar articulados entre si.
2 – Aperfeiçoamento de docentes para o ensino superior	O ensino superior é um setor de formação de recursos humanos para os demais níveis de ensino e para a sociedade; os cursos de Pós-Graduação no sentido estrito – Mestrado e Doutorado – devem ser regularmente dirigidos para a formação de recursos humanos para o próprio ensino superior.
3 – Internacionalização	A capacitação dos docentes das instituições brasileiras deve ser programada em função das capacidades de atendimento dos cursos aqui localizados; nos casos específicos de impossibilidade de atendimento em âmbito nacional, devem ser programados convênios e intercâmbios com instituições estrangeiras.

Fonte: Adaptado de MEC (1975).

Pelo Quadro 1, entende-se que o PNPG, em sua primeira edição, de 1975, teve o propósito de oportunizar o desenvolvimento da Pós-Graduação como modalidade oficial de ensino e de complementação aos demais níveis, com recursos oriundos do poder público para a sua oficialização e funcionamento (Ortega; Brandão, 2020). Além disso, por meio do I PNPG, foi implementado o Sistema de Pós-Graduação,

[...] formado por um conjunto de universidades e outras instituições de ensino superior, com propostas e perfis diversos, com autonomia para criar ou fechar cursos, mas dependentes do financiamento da Capes e do seu *ranking*, que lhes atribui uma classificação. Cabe a esse sistema o atendimento das demandas de formação de professores e pesquisadores, voltados para atividades de ensino e pesquisa e para a necessidade de formar técnicos altamente especializados para os setores público e privado (Nobre; Freitas, 2017, p. 27-28).

A partir de 1980, cinco novos PNPG foram criados, com novas diretrizes. Ortega e Brandão (2020) ressaltam que, em cada um dos PNPG, novas medidas foram sugeridas, muitas delas não se concretizando, fato esse que demonstra os desafios existentes no desenvolvimento da pesquisa brasileira. As principais mudanças que caracterizaram cada um dos planos e seus respectivos anos de adesão são:

- **II Plano (1982 a 1985):** durante o governo do militar João Baptista de Oliveira Figueiredo, focou na qualidade do ensino superior, centrando suas ações em maior participação da comunidade;
- **III Plano (1986 a 1989):** durante o governo de José Sarney de Araújo Costa, enfocou preencher as lacunas dos planos anteriores e aproximar a Pós-Graduação da universidade, sobretudo para a formação dos docentes;
- **IV Plano (1994 a 2002):** durante os governos de Itamar Augusto Cautiero Franco e Fernando Henrique Cardoso, houve discussão sobre financiamentos; elaboração de plano articulado entre universidade e programas; processo de avaliação pela Capes; capacitação dos docentes; aprimoramento constante da qualidade dos programas, dentre outras ações (Ortega; Brandão, 2020);
- **V Plano (2005 a 2010):** durante o governo de Luiz Inácio Lula da Silva, enfoque na criação de novos programas, conforme as demandas do mercado e alinhamento do que é ensinado e pesquisado com as necessidades da sociedade com incentivo do Mestrado Profissional (Barbosa, 2020);
- **VI Plano (2011-2020):** durante os governos de Dilma Vana Rousseff, Michel Miguel Elias Temer Lulia e Jair Messias Bolsonaro, foco na criação de programas fundamentados na interdisciplinaridade, com flexibilidade curricular, em conformidade com as demandas do mercado (Barbosa, 2020).

Importante esclarecer que no V Plano, uma nova modalidade de curso foi mencionada: o Mestrado Profissional. Essa modalidade já havia sido regulamentada no ano de 1995, por meio da Portaria n. 47, de 17 de outubro de 1995. A proposta desse curso é que a Pós-Graduação ofereça formação de recursos humanos voltados para as empresas e que contribuam diretamente para a inovação científico-tecnológica no País. Do mesmo modo, no ano de 2017, instituiu-se o Doutorado Profissional, por meio da Portaria n. 389, de 23 de março de 2017, do MEC (Barbosa, 2020).

Os seis planos foram concebidos como os principais impulsionadores no avanço da Pós-Graduação brasileira, ao apresentar normas de regulamentação e, sobretudo, parâmetros para garantia da qualidade dos resultados das pesquisas, visando aos impactos dos cursos na sociedade (Luna; Luna, 2015). Embora com a existência de muitos desafios, percebe-se que “[...] o investimento brasileiro em ciência e tecnologia tem aumentado significativamente e apresenta várias vertentes em seu financiamento, o que fez com que a Pós-Graduação do País evoluísse [...]” (Luna; Luna, 2015, p. 1).

A expansão das universidades e os altos investimentos do governo no ensino superior foram essenciais para garantir que muitas das propostas almejadas nos planos fossem alcançadas. Exemplo de resultados desse processo foi o aumento do número de cursos, em diversas áreas do conhecimento (Luna; Luna, 2015; Nobre; Freitas, 2017). Conforme dados publicados na plataforma do Sistema de Informações Georreferenciadas da Capes (Geocapes, 2022), o Brasil possui um total de 4.559 Programas de Pós-Graduação. A distribuição dos programas por grau (modalidade dos cursos) é apresentada no Quadro 2.

Quadro 2 – Quantidade de programas de Pós-Graduação no Brasil por grau em 2020

Grau – modalidade dos cursos	Quantidade de programas
Mestrado e Doutorado	2.356
Mestrado	1.282
Mestrado Profissional	800
Doutorado	78
Mestrado Profissional e Doutorado Profissional	41
Doutorado Profissional	2
Total	4.559

Fonte: Adaptado de Geocapes (2022). Nota: Os valores referem-se ao número de programas conforme categorização do GEOCAPES em 2020. Programas que oferecem mais de um grau foram considerados uma única unidade.

Observa-se no quadro 2 apenas dois programas de doutorado profissional, o que pode ser explicado pelo fato de a modalidade de doutorado profissional haver sido aprovada na Portaria nº 389, de 23 de março de 2017 (Brasil, 2017b), e, dessa forma, os cursos começam a ser recomendados em 2018, com início das atividades para 2019.

Além da quantidade expressiva do número de programas, observa-se que todas as nove grandes áreas do conhecimento estabelecidas pela Capes são contempladas, conforme dados extraídos do Geocapes (2022), expostos no Quadro 3.

Quadro 3 – Quantidade de programas de Pós-Graduação por grandes áreas: ano 2020

Grande área	DOA	MEA	MEA/ DOA	MEA/ DOA/ MEP	MEA/ MEP	MEP DOP	MEP	DOP	Total por área
Ciências agrárias	1	111	269	-	-	1	43	-	425
Ciências biológicas	3	61	231	-	-	1	18	-	314
Ciências da saúde	23	136	407	-	-	7	148	-	721
Ciências exatas e da terra	10	97	210	-	-	1	30	-	348
Ciências humanas	3	192	365	-	-	5	100	-	665
Ciências sociais aplicadas	5	205	266	-	-	8	147	-	631
Engenharias	7	149	212	-	-	1	75	-	444
Linguística, letras e artes	1	67	138	-	-	-	19	-	225
Multidisciplinar	25	264	258		-	17	220	2	786
Total:	78	1.282	2.356	-	-	41	800	2	4.559

Fonte: geocapes.capes.gov.br/geocapesds, extraído em 16/11/2021.

Nota: DOA: Doutorado acadêmico

MEA: Mestrado acadêmico

MEA/DOA: Mestrado acadêmico e Doutorado acadêmico

MEA/DOA/MEP: Mestrado acadêmico e doutorado acadêmico e mestrado profissional

MEA/MEP: Mestrado acadêmico e mestrado profissional

MEA/DOP: Mestrado profissional e Doutorado profissional

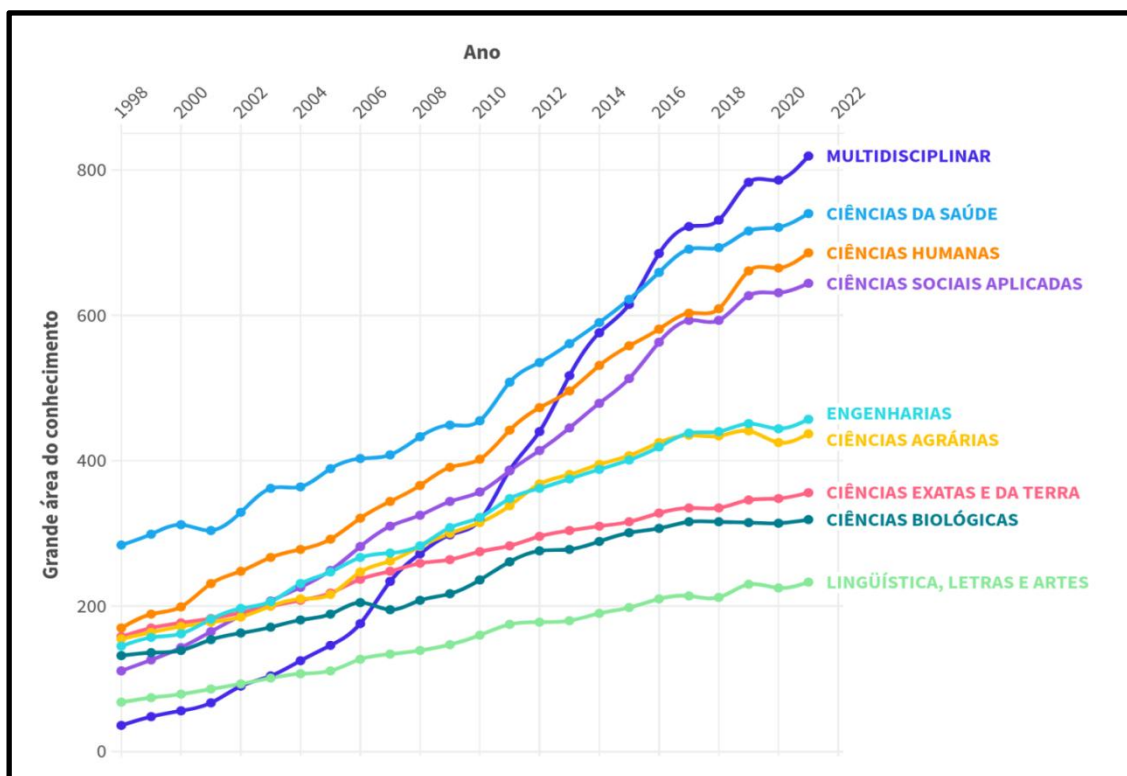
MEP: Mestrado profissional

DOP: Doutorado profissional

No ano de 2020, a área com maior número de programas é a área Multidisciplinar, com 786 programas, além de ser a única área que, até o momento, apresenta Doutorado Profissional.

Ao comparar o número de programas por áreas de conhecimento, de 1998 a 2020, é possível identificar, também, um crescimento exponencial. O crescimento do número de programas por grandes áreas é exposto no Gráfico 1.

Gráfico 1 – Crescimento do número de PPG por grande área: 1998 a 2020



Fonte: geocapes.capes.gov.br/geocapesds, extraído em 09/08/2023.

O crescimento observado no Gráfico 1 é justificado pela vinculação crescente das políticas educacionais com a política de ciência e tecnologia promovida pelo Estado brasileiro “[...] e que passa pela criação de uma infraestrutura de ensino e pesquisa que impulse a produção e utilização do conhecimento científico e tecnológico no País [...]” (Barbosa, 2020, p. 2). Além disso, o trabalho realizado pela Capes e pelo CNPq foi fundamental para a alocação de recursos, por conseguinte, promoveu a expansão dos cursos em todas as áreas do conhecimento, sob um enfoque interdisciplinar, congregando pesquisadores de diferentes áreas no desdobramento de novas descobertas (Barbosa, 2020).

Outro dado expressivo está no número de discentes, que cresceu ao longo do tempo, sobretudo com a criação de novos programas. De acordo com dados armazenados no Geocapes, o Brasil possui um total de 304.892 pós-graduandos matriculados e 80.105 egressos.

O Quadro 4 apresenta a distribuição dos discentes por grau e por situação (matriculada ou egressa).

Quadro 4 – Quantidade de discentes por grau e por situação, ano base 2020

Grau – modalidade dos cursos	Matriculados	%	Egressos – Titulados	%
Mestrado:	136.194	44.67%	46.060	57.50%
Doutorado:	124.530	40.84%	20.066	25.05%
Mestrado Profissional:	44.168	14.49%	13.979	17.45%
Doutorado Profissional:	755	0.25%	9	0.01%
Total:	304.892		80.105	

Fonte: geocapes.capes.gov.br/geocapesds, extraído em 12/11/2021.

Pelos dados descritos no Quadro 4, o crescimento do número de alunos está associado, dentre outros fatores, à expansão do número de vagas e à criação de novos cursos, processo esse ocorrido em paralelo com a reformulação das universidades (Luna; Luna, 2015). Como já dito, as atividades da modalidade de doutorado profissional iniciaram em 2019, e os primeiros alunos titulados em 2020 (Audy, 2021). Além disso, o resultado está em consonância com a meta projetada no VI Plano, ao discorrer que é prioridade da Pós-Graduação para a segunda década do século XXI apoiar a educação básica e outros níveis e modalidades de ensino, bem como a criar programas de Pós-Graduação em formatos inovadores que ofereçam maior interdisciplinaridade e flexibilidade curricular (Barbosa, 2020).

No percurso histórico da Pós-Graduação brasileira, é preciso compreender as transformações institucionais e temáticas que marcaram sua consolidação ao longo das décadas. Ao explorar o desenvolvimento desse nível educacional no Brasil, Almeida (2017) propõe uma periodização analítica composta por três grandes fases, caracterizadas por aspectos pedagógicos, estruturais, avaliativos e político-institucionais. Essa proposta permite interpretar a Pós-Graduação não apenas como um arranjo formal, mas como um campo em constante reconfiguração, marcado por diferentes agendas e influências. O Quadro 5 apresenta essas três fases, articulando os contextos históricos, as características predominantes e os períodos correspondentes.

Quadro 5 – Periodização analítica da Pós-Graduação brasileira segundo Almeida (2017)

Fase	Síntese	Período
1) Modernização educacional e surgimento da Pós-Graduação	A Pós-Graduação consolida-se a partir das propostas do educador Anísio Teixeira, que defendia a integração entre ensino e pesquisa como estratégia de modernização educacional. Sua atuação como gestor público e defensor da institucionalização da pesquisa educacional foi central para o desenvolvimento inicial da Pós no Brasil.	De 1960 a 1970

Fase	Síntese	Período
2) Fragilidades teórico-metodológicas na produção pós-graduada	Esse período é marcado pelo avanço na criação de cursos stricto sensu, mas também pela identificação de fragilidades na produção científica, atribuídas à adesão a modismos teóricos, à formação insuficiente dos pesquisadores e às limitações estruturais dos programas.	De 1970 a 1990
3) Regulação estatal e internacionalização da Pós-Graduação	A partir dos anos 1990, intensificam-se as políticas de avaliação e consolidação de parâmetros produtivistas por meio dos Planos Nacionais de Pós-Graduação. Destaca-se também o debate sobre internacionalização e os efeitos do Tratado de Bolonha no Brasil, com implicações na estrutura e nos objetivos da formação pós-graduada.	Desde meados de 1990 até os dias atuais, considerando que o estudo foi feito em 2017

Fonte: Adaptado de Almeida (2017).

De acordo com o Quadro 5, a fase mais recente da Pós-Graduação brasileira é marcada pela intensificação das políticas de avaliação, pelo fortalecimento de parâmetros produtivistas e pela crescente internacionalização, em especial após os anos 1990 (Almeida, 2017).

É preciso destacar que nos anos recentes, a CAPES lançou a iniciativa de Governança de Dados da Pós-Graduação (GoPG), que visa estabelecer princípios, políticas e práticas para garantir a qualidade, integridade e interoperabilidade das informações que sustentam o sistema nacional de pós-graduação, conforme apresentado no Relatório Imersivo Sucupira 2.0 (Capes, 2023). Instituído formalmente em 2022, o GoPG se articula como um esforço estratégico da CAPES para consolidar uma gestão mais eficiente dos dados coletados pela Plataforma Sucupira, promovendo maior transparência, confiabilidade e suporte à formulação de políticas públicas e processos avaliativos. Embora represente um avanço institucional significativo, trata-se de uma ação recente, cuja implementação ainda está em curso e cujos efeitos concretos na melhoria dos dados disponibilizados demandam tempo para serem plenamente consolidados.

No contexto desta pesquisa, o GoPG se configura como um elemento promissor, uma vez que os dados abertos da CAPES são insumos importantes em para diversas investigações. Nesse contexto, os estudos bibliométricos tornam-se particularmente úteis, pois permitem quantificar a produção científica dos programas, evidenciar indicadores de desempenho e subsidiar a avaliação das atividades acadêmicas. Como destacam Nascimento, Pinto e Dias (2020), tais estudos constituem uma estratégia metodológica eficaz para mensurar o conhecimento produzido, prospectar tendências e contribuir para o aprimoramento da própria Pós-Graduação.

Com efeito, ao aplicar estudos que quantifiquem a produção científica, dentre outras ações relacionadas à pesquisa, torna-se possível analisar “[...] os conhecimentos produzidos

gerando indicadores utilizados para mensurá-los, proporcionando para ciência a verificação do que está sendo produzido, prospectando e contribuindo para o avanço do conhecimento” (Nascimento; Pinto; Dias, 2020, p. 209). De fato, os estudos bibliométricos, conforme pontuado por esses autores, são variados e podem ser utilizados em qualquer área de conhecimento, sobretudo na Ciência da Informação, visto que é nessa área que eles foram desenvolvidos e são aperfeiçoados.

2.3 A Ciência da Informação no Brasil

A Ciência da Informação pode ser definida como uma área do conhecimento científico que se preocupa com o estudo da informação e o seu uso para a geração de conhecimento em diferentes contextos (Pinto, 2018). A clássica definição de Harold Borko revela a abrangência da área quando se trabalha com a informação e o seu dinamismo. Assim, a Ciência da Informação “[...] está relacionado ao corpo de conhecimento relacionado à origem, coleta, organização, armazenamento, recuperação, interpretação, transmissão, transformação e uso da informação [...]” (Borko, 1968, p. 3, tradução nossa).

A Ciência da Informação nasceu após a Segunda Guerra Mundial, assim como vários outros campos, tendo como foco dar solução ao “[...] problema da explosão de informações e usando a tecnologia como solução” e percebe-se, na contemporaneidade, que “a ciência da informação lida com os mesmos problemas na Web e em ambientes digitais [...]” (Bates; Maack, 2009, p. 1, tradução nossa).

A gênese da Ciência da Informação remete às décadas de 1950 e 1960, sendo que a explosão informacional, a sua interação e relação com outras áreas do conhecimento e o avanço tecnológico constituem as três causas principais do nascimento desse campo (Saracevic, 1996). Dentre alguns marcos históricos mundiais e oficiais de institucionalização dessa ciência, citam-se: a Conferência Internacional sobre Informação Científica, realizada em Washington, de 16 a 21 de novembro de 1958 (Bates; Maack, 2009) e a mudança de nome do Instituto Americano de Documentação para Sociedade Americana de Ciência da Informação, no início da década de 1960 (Borko, 1968).

Assim, como área do conhecimento institucionalizada, a Ciência da Informação tem origem na década de 1960, liderada pelos Estados Unidos, numa visão estratégica e política de informação, tendo uma estreita relação com C&T, “tanto que a chamada informação científica e tecnológica praticamente dá origem à área” (Pinheiro, 2000, p. 3). Para resolver os problemas relativos à informação, suas variadas formas, os fluxos e processos que permeiam a

conversão da informação para o seu uso pela sociedade, a Ciência da Informação recorre a outras disciplinas, evidenciando uma estreita relação com “a Comunicação, a Biblioteconomia e a Ciência da Computação” (Pinheiro, 2000, p. 9).

Ainda que se possa estabelecer que existe relações intensas e profundas entre a Ciência da Informação e outras áreas do conhecimento, como destaca a autora, pois “na fértil literatura teórica estrangeira sobre a questão há o reconhecimento do diálogo interdisciplinar”, também se encontra nessa mesma literatura a ressalva de que não há a “fusão de áreas” (Pinheiro, 2000, p. 9). Sobre isso, Saracevic (1996) apontou a origem do campo, relatando a sua identidade, que foi formada a partir da interação com outras áreas do conhecimento, em específico a Biblioteconomia, Ciência da Computação, Ciência Cognitiva (incluindo Inteligência Artificial) e a Comunicação, sem, contudo, determinar que houve a fusão delas.

No Brasil, a Ciência da Informação teve origem surgiu na década de 1970, a partir de diversas iniciativas necessárias para a consolidação de um campo científico, tais como: criação de associações de profissionais, de institutos de pesquisa, de periódicos científicos e de programas de Pós-Graduação (Araújo; Valentim, 2019). Nessa década, ocorreu a implementação do IBICT, antigo Instituto Brasileiro de Bibliografia e Documentação (IBBD), como também a criação do periódico Ciência da Informação e do Programa de Pós-Graduação em Ciência da Informação, ambos gerenciados pelo Instituto (Marteletto, 2009). Devido ao contexto histórico, sociopolítico e cultural do Brasil, a sua implementação no país “permitiu a construção de um cenário informacional distinto e particular, em que pese a universalidade da informação” (Pinheiro, 2000, p. 3). Segundo a autora, se “nos EUA as atividades de informação estão concentradas no setor privado, no Brasil ocorre o inverso”, ainda que nos Estados Unidos as “atividades de informação científica e técnica são estimuladas, promovidas e financiadas pelo Estado” (Pinheiro, 2000, p. 3).

De acordo com Justino (2019), a criação do IBICT foi fundamental para alavancar o interesse de pesquisa na área da informação, que se desenvolveu de modo paralelo às escolas de Biblioteconomia e Documentação, fato esse que confirma a forte relação da Ciência da Informação com essas áreas do conhecimento. A partir da década de 1970, novos programas de Pós-Graduação e periódicos em Ciência da Informação foram criados, cuja pesquisa nessa área foi fortalecida, abrangendo pesquisadores de diferentes campos do conhecimento e, assim, mantendo o aspecto inter-relacional da área, tal como ocorria em países europeus e nos Estados Unidos (Marteletto, 2009; Justino, 2019).

Araújo e Valentim (2019) consideram que a Ciência da Informação brasileira replicou os modelos de pesquisa de países como os Estados Unidos e países da Europa e da América

Latina, com uma pequena participação de outros países, cujas aplicações práticas das pesquisas se voltaram para a realidade brasileira, o que evidencia uma agenda própria de pesquisa. Assim, pesquisadores brasileiros “[...] utilizam-se de teorias, conceitos e métodos de pesquisa criados e usados nestes países, os aplicando, os misturando, mas também os reelaborando em face dos problemas de pesquisa nacionais [...]” (Araújo; Valentim, 2019, p. 233). Além disso, o Brasil tem se destacado “[...] na criação de modelos, teorias e conceitos próprios, alguns dos quais também apresentados e discutidos internacionalmente” (Araújo; Valentim, 2019, p. 233).

Silva (2011) relata que, em termos epistemológicos, a Ciência da Informação brasileira reflete os fundamentos oriundos de países europeus e dos Estados Unidos e que, em termos históricos, o campo científico se desenvolveu sob dois aspectos: um de natureza institucional ou governamental e outro de natureza acadêmica. O primeiro aspecto está relacionado à criação do IBICT e que evidencia o compromisso social e profissional da área, com investimentos do governo na formação de profissionais para atender as demandas de mercado. O segundo aspecto, consequência do primeiro, diz respeito aos investimentos em pesquisas, mediante a criação dos cursos de Pós-Graduação, a maioria oferecida junto aos cursos de Graduação em Biblioteconomia e Arquivologia (Silva, 2011).

A criação do IBBD e depois IBICT representou um marco histórico e governamental para a consolidação da Ciência da Informação como campo de pesquisa no contexto brasileiro. Isso porque, segundo Silva (2011), foi reconhecida a necessidade de políticas públicas na oferta de cursos e vagas para formação de profissionais nas universidades, sobretudo bibliotecários, e o desenvolvimento de pesquisas que garantissem novas descobertas no que tange à informação e à tecnologia. A Ciência da Informação, pesquisando com insumos e em conjunto com áreas como Administração, Arquivologia, Biblioteconomia, Ciência da Computação, dentre outras, foi uma proposta fértil para resolver inúmeros problemas econômicos e estimular o desenvolvimento do País (Lança; Amaral; Gracioso, 2018).

A evolução de um campo do conhecimento, [...] vai ser determinada e delimitada por essas relações interdisciplinares [ou multidisciplinares], em processo de constante mutação, como um organismo vivo. As transmutações são determinadas por fenômenos científicos, tecnológicos, econômicos, históricos, sociais e culturais (Lança; Amaral; Gracioso, 2018, p. 159).

Essas transmutações promoveram, sobretudo nas últimas décadas do século XX, mudanças culturais, políticas e sociais em todos os países, incluindo o Brasil. A Ciência da Informação nasce, no País, a partir desse contexto, e foi estruturada, em essência, pela criação de associações profissionais preocupadas em legitimar o campo e, ao mesmo tempo, formar

mão de obra habilitada para atuação na sociedade (Marteleto, 2009). É constituída, assim, a Associação Brasileira de Educação em Ciência da Informação (ABECIN), voltada à formação e à pesquisa em nível de Graduação, e a Associação Nacional de Pesquisa e Pós-Graduação em Ciência da Informação (ANCIB), cuja abrangência contempla a formação e a pesquisa em nível de Pós-Graduação (Araújo; Valentim, 2019).

A Abecin é referência na estruturação e funcionamento dos cursos oferecidos para a formação inicial de profissionais na área da informação, pois apresenta diretrizes e conteúdos suficientes para garantir uma educação integral e de qualidade, em consonância com as demandas do mercado (Barbalho *et al.*, 2021). De modo complementar, a ANCIB tem como finalidade acompanhar e estimular as atividades de pesquisa desenvolvidas na área, e isso faz com que a associação se constitua “[...] uma importante instância de representação científica e política das questões inerentes à área de informação” (Araújo; Valentim, 2019, p. 234).

Dentre as diversas iniciativas promovidas pela Ancib para o fortalecimento da Ciência da Informação, além da estruturação dos PPG, tem-se o planejamento e execução de um dos maiores eventos nacionais da área, o Encontro Nacional de Pesquisa e Pós-Graduação em Ciência da Informação (ENANCIB) (Silva, 2019). Esse evento é referência nacional e se apresenta como “[...] uma das mais importantes vitrines que permitem a visualização das pesquisas que vêm sendo desenvolvidas especialmente nos cursos de Pós-Graduação em Ciência da Informação acerca das questões epistemológicas e históricas dessa área” (Silva, 2019, p. 60).

A Ancib tem trabalhado com a gestão de Grupos de Trabalho (GT) que tentam representar o panorama temático das pesquisas desenvolvidas no Brasil. Esses grupos constituem a espinha dorsal de organização do Enancib e são coordenados por pesquisadores de diversas instituições brasileiras, de modo a acompanhar a realidade de todas as regiões brasileiras (Ancib, 2022). O Quadro 6 apresenta os 12 GT e suas respectivas ementas.

Quadro 6 – Grupos de Trabalho da Ancib e suas ementas

Grupo de Trabalho	Ementas
I - Estudos Históricos e Epistemológicos da Informação	Estudos históricos e epistemológicos da Ciência da Informação (escolas de pensamento, correntes teóricas, autores e obras de fundamentação, leituras teórico-metodológicas e conceituações). Constituição, desenvolvimento e inovação conceitual, teórica e metodológica do campo científico informacional. Os objetos de estudos da Ciência da Informação e suas transformações teórico-conceituais. Reflexões e discussões sobre disciplinaridade, interdisciplinaridade e transdisciplinaridade.
II - Organização e Representação do Conhecimento	Teorias, metodologias, políticas, instrumentos, processos e produtos para a organização e representação do conhecimento, recuperação e acesso à informação, nas suas dimensões epistemológicas, aplicadas, sociais, culturais e terminológicas

Grupo de Trabalho	Ementas
	enquanto conhecimento socializado, institucionalizado ou não, em ambientes informacionais (tais como: arquivos, museus, bibliotecas e congêneres), incluindo o uso e desenvolvimento das tecnologias de informação e as relações inter, multi e transdisciplinares neles verificadas.
III - Mediação, Circulação e Apropriação da Informação	Estudo dos processos e das relações entre mediação, circulação e apropriação de informações, em diferentes contextos e tempos históricos, considerados em sua complexidade, dinamismo e abrangência, bem como relacionados à construção e ao avanço do campo científico da Ciência da Informação, compreendido em dimensões inter e transdisciplinares, envolvendo múltiplos saberes e temáticas, bem como contribuições teórico-metodológicas diversificadas em sua constituição.
IV - Gestão da Informação do Conhecimento	Gestão de ambientes, sistemas, unidades, serviços, produtos de informação e recursos informacionais. Estudos de fluxos, processos, usos e usuários da informação como instrumentos de gestão. Gestão do conhecimento e aprendizagem organizacional no contexto da Ciência da Informação. Marketing da informação, monitoramento ambiental e inteligência competitiva. Estudos de redes para a gestão. Aplicação das tecnologias de informação e comunicação à gestão da informação e do conhecimento.
V - Política e Economia da Informação	Políticas e regimes de informação. Informação, Estado e governo. Propriedade intelectual. Acesso à informação. Economia política da informação e da comunicação. Produção colaborativa. Poder, ativismo e cidadania. Conhecimento, aprendizagem e inovação. Ética da informação. Informação e ecologia.
VI - Informação, Educação e Trabalho	O mundo do trabalho informacional: atores, cenários, competência em informação, dimensões e habilidades. Organização, processos de trabalho em dispositivos de informação e cultura. As relações entre informação, educação, trabalho, saúde e tecnologia. Regulamentação profissional, entidades sindicais, associações de classe e mercado de trabalho e competência profissional. Diversidade cultural, representações sociais, práticas e construção identitária dos profissionais da informação. Responsabilidade social, ética e profissional na Ciência da Informação. As bases curriculares e experiências pedagógicas: formação e perfil profissional ou docente.
VII - Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação	Estudos teóricos, aplicados e metodológicos sobre a produção, comunicação e uso da informação em Ciência, Tecnologia e Inovação. Inclui pesquisas relacionadas aos processos de comunicação, divulgação, análise e formulação de indicadores para planejamento, avaliação e gestão em Ciência e Tecnologia.
VIII - Informação e Tecnologia	Estudos e pesquisas teórico-práticos sobre e para o desenvolvimento de tecnologias de informação e comunicação que envolvam os processos de geração, representação, armazenamento, recuperação, disseminação, uso, gestão, segurança e preservação da informação em ambientes digitais.
XIX - Museu, Patrimônio e Informação	Análise das relações entre o museu (fenômeno cultural), o patrimônio (valor simbólico) e a informação (processo), sob múltiplas perspectivas teóricas e práticas de análise. Museu, patrimônio e informação: interações e representações. Patrimônio musealizado: aspectos informacionais e comunicacionais.
X - Informação e Memória	Estudos sobre a relação entre os campos de conhecimento da Ciência da Informação e da Memória Social. Pesquisas transdisciplinares que envolvem conceitos, teorias e práticas do binômio 'informação e memória'. Memória coletiva, coleções e colecionismo, discurso e memória. Representações sociais e conhecimento. Articulação entre arte, cultura, tecnologia, informação e memória, por meio de seus referenciais, na contemporaneidade. Preservação e virtualização da memória social.
XI - Informação e Saúde	Estudos das teorias, métodos, estruturas e processos informacionais em diferentes contextos da saúde, considerada em sua abrangência e complexidade. Impacto da informação, tecnologias e inovação em saúde. Informação nas organizações de saúde. Informação, saúde e sociedade. Políticas de informação em saúde. Formação e capacitação em informação em saúde.

Grupo de Trabalho	Ementas
XII - Informação, Estudos Étnico-Raciais, Gênero e Diversidades	Estudos teóricos e aplicados em informação sobre Raça, Classe, Gênero, Sexualidades e Interseccionalidades. Teorias Críticas, Culturais, Racial, Feministas e Queer. Correntes teóricas, escolas de pensamento, bases metodológicas-conceituais e aplicações técnico-científicas dos estudos étnico-raciais, de gênero e de diversidade. Teorias, discursos, saberes, atividades científicas e profissionais em ambientes informacionais comunitários, populares e organizacionais. Relações sociais, de poder e resistências. Epistemicídio, violências e insurgências. Estudos Pós-Coloniais, Decoloniais e Anticoloniais. Estudos Críticos da Branquitude. Justiça Social, Informacional, Racial e de Gênero.

Fonte: Adaptado de Silva (2019) e Ancib (2022).

A análise às ementas dos GT evidencia o amplo escopo de pesquisa da Ciência da Informação, pesquisas essas desenvolvidas, na maioria dos casos, junto aos PPG da área (Lança; Amaral; Gracioso, 2018; Silva, 2019). De acordo com a pesquisa de Lança, Amaral e Gracioso (2018), o Brasil possui um número considerável de PPG em Ciência da Informação, incluindo programas de natureza acadêmica e profissional. Esses autores concluíram que a atividade de pesquisa dos PPG revela a interação de diversas áreas do conhecimento na atividade científica da área, o que, conseqüentemente, garante o seu fortalecimento e a sua visibilidade como área do conhecimento no Sistema Nacional de Pós-Graduação. O Quadro 7 elenca os 27 programas em vigência, distribuídos em 24 instituições, os locais a que pertencem e o conceito atribuído pela avaliação realizada pela Capes⁴.

Quadro 7 – Programas de Pós-Graduação em Ciência da Informação no Brasil – 2022

Instituição	UF	Programas	M E	D O	M P	D P
FUNDAÇÃO CASA DE RUI BARBOSA (FCRB)	RJ	MEMÓRIA E ACERVOS	-	-	3	-
FUNDAÇÃO UNIVERSIDADE FEDERAL DE SERGIPE (FUFSE)	SE	CIÊNCIA DA INFORMAÇÃO	-	-	3	-
UNIVERSIDADE DE BRASÍLIA (UNB)	DF	CIÊNCIAS DA INFORMAÇÃO	5	5	-	-
UNIVERSIDADE DE SÃO PAULO (USP)	SP	CIÊNCIA DA INFORMAÇÃO	4	4	-	-
		GESTÃO DA INFORMAÇÃO	-	-	4	-
UNIVERSIDADE DO ESTADO DE SANTA CATARINA (UDESC)	SC	GESTÃO DA INFORMAÇÃO	-	-	3	-
UNIVERSIDADE ESTADUAL DE LONDRINA (UEL)	PR	CIÊNCIA DA INFORMAÇÃO	4	4	-	-
UNIVERSIDADE ESTADUAL PAULISTA JÚLIO DE MESQUITA FILHO, MARÍLIA (UNESP-MAR)	SP	CIÊNCIA DA INFORMAÇÃO	6	6	-	-
UNIVERSIDADE FEDERAL DA BAHIA (UFBA)	BA	CIÊNCIA DA INFORMAÇÃO	4	4	-	-
UNIVERSIDADE FEDERAL DA PARAÍBA, JOÃO PESSOA (UFPB-JP)	PB	CIÊNCIA DA INFORMAÇÃO	4	4	-	-

⁴ A Capes, por meio do Sistema de Avaliação da Pós-Graduação, atribui uma nota, de 1 a 7, considerando as atividades desenvolvidas nos programas. Se a nota obtida for superior a 2, isso indica a legitimação e reconhecimento, por conseguinte, permite o funcionamento do programa (Capes, 2022c).

Instituição	UF	Programas	M E	D O	M P	D P
UNIVERSIDADE FEDERAL DE ALAGOAS (UFAL)	AL	CIÊNCIA DA INFORMAÇÃO	A	-	-	-
UNIVERSIDADE FEDERAL DE MINAS GERAIS (UFMG)	MG	CIÊNCIAS DA INFORMAÇÃO	5	5	-	-
		GESTÃO & ORGANIZAÇÃO DO CONHECIMENTO	5	5	-	-
UNIVERSIDADE FEDERAL DE PERNAMBUCO (UFPE)	PE	CIÊNCIA DA INFORMAÇÃO	4	4	-	-
UNIVERSIDADE FEDERAL DE SANTA CATARINA (UFSC)	SC	CIÊNCIA DA INFORMAÇÃO	5	5	-	-
UNIVERSIDADE FEDERAL DE SÃO CARLOS (UFSCAR)	SP	CIÊNCIA DA INFORMAÇÃO	3	-	-	-
UNIVERSIDADE FEDERAL DO CARIRI (UFCA)	CE	BIBLIOTECONOMIA	-	-	3	-
UNIVERSIDADE FEDERAL DO CEARÁ (UFC)	CE	CIÊNCIA DA INFORMAÇÃO	3	-	-	-
UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO (UFES)	ES	CIÊNCIA DA INFORMAÇÃO	A	-	-	-
UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO (UNIRIO)	RJ	BIBLIOTECONOMIA	-	-	3	-
		GESTÃO DE DOCUMENTOS E ARQUIVOS	-	-	3	-
UNIVERSIDADE FEDERAL DO PARÁ (UFPA)	PA	CIÊNCIA DA INFORMAÇÃO	3	-	-	-
UNIVERSIDADE FEDERAL DO RIO DE JANEIRO (UFRJ)	RJ	CIÊNCIA DA INFORMAÇÃO - UFRJ - IBICT	4	4	-	-
UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE (UFRN)	RN	GESTÃO DA INFORMAÇÃO E DO CONHECIMENTO	-	-	3	-
UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL (UFRGS)	RS	CIÊNCIA DA INFORMAÇÃO	A	-	-	-
UNIVERSIDADE FEDERAL FLUMINENSE (UFF)	RJ	CIÊNCIA DA INFORMAÇÃO	4	4	-	-
UNIVERSIDADE FUMEC (FUMEC)	MG	SISTEMAS DE INFORMAÇÃO E GESTÃO DO CONHECIMENTO	4	4	-	-

Fonte: Adaptado de Lança, Amaral e Gracioso (2018) e da Plataforma Sucupira (Capes, 2022d).

Legenda: ME: Mestrado Acadêmico; DO: Doutorado Acadêmico; MP: Mestrado Profissional; DP: Doutorado Profissional

Os PPG em Ciência da Informação, conforme descrito no Quadro 7, são fundamentais para a consolidação da área no Brasil, pois são as pesquisas que irão gerar conhecimento novo necessário para o avanço da sociedade. Tais programas representam boa parte da memória da Ciência da Informação, constituindo-se, assim, como laboratórios de descobertas, que formam os novos pesquisadores (mestres e doutores), que são absorvidos pelo próprio sistema científico do País (Oliveira, 2021).

Ao analisar a trajetória histórica dos PPG em Ciência da Informação, Lança, Amaral e Gracioso (2018) identificaram que o Programa mais antigo é o da UFRJ, que implementou o primeiro Mestrado em Ciência da Informação do Brasil, fundado em 1970 pelo IBICT. As autoras também relataram que três programas se destacam por terem maior tempo de

existência, tal como o Programa Pós-Graduação em Ciência da Informação⁵ (PPGCI) da Universidade Federal de Minas Gerais (UFMG), iniciado em 1976 com mestrado e que agregou o doutorado em 1997; o da Universidade de Brasília (UnB), implementado em 1978; e o da Universidade Estadual Paulista (Unesp), criado no ano de 1988. No que tange à quantidade de programas profissionais, neste ano de 2022 existem oito programas profissionais e essa ocorrência pode estar associada ao fato de que a Ciência da Informação possui uma característica de aplicação, ou seja, relaciona-se com outras áreas de conhecimento, fomentando a interdisciplinaridade e beneficiando a sociedade (Lança; Amaral; Gracioso, 2018)

A partir dos resultados das pesquisas desenvolvidas nos PPG e de outras atividades inerentes ao ensino, pesquisa e ao trabalho profissional na área da informação, foram criados os periódicos, caracterizados como canais formais que comunicam os resultados das pesquisas e das experiências (Campello; Cendón; Kremer, 2003), já citados na seção 3.1. Ao longo da história da Ciência da Informação brasileira, nota-se a existência de periódicos que vêm acompanhando as novidades do mundo digital, como migração para o formato digital, divulgação e citações em redes sociais (altimetria) e menção nos gerenciadores de referências (Borba; Caregnato, 2021).

Assim como Borba e Caregnato (2021), Lunardelli, Paiva e Kawakami (2020) também demonstraram o potencial dos periódicos da Ciência da Informação, inclusive, analisando os elementos de padronização dos periódicos, como os resumos. Mediante a identificação de 24 periódicos, foi constatado que há um imperativo dessa área em manter o aprimoramento de diretrizes, com a finalidade de comunicar os resultados das pesquisas, de modo padronizado, mantendo-se, portanto, um nível de confiança e qualidade no que é disseminado (Lunardelli; Paiva; Kawakami, 2020). As publicações nos periódicos, segundo Araújo e Miguel (2017), constituem o principal produto proveniente do trabalho dos pesquisadores, cujas descobertas são disseminadas e aplicadas pelos profissionais.

⁵ Na cronologia do programa PPGCI da UFMG, destaca-se que em 1976 houve o início do Mestrado em Biblioteconomia, e somente em 1991 houve a mudança da sua denominação para Curso de Programa de Pós-Graduação em Ciência da Informação, iniciado ainda somente com o curso de Mestrado. Em 1997, inicia-se o oferecimento do curso de Doutorado, cuja criação se deu em aprovação em julho de 1996, pela UFMG e a CAPES.

2.3.1 A Ciência da Informação como ciência social aplicada: programas e cursos de pós-graduação

A compreensão da Ciência da Informação como ciência social aplicada requer uma abordagem epistemológica que vá além das concepções técnico-instrumentais e incorpore o papel social da informação, dos documentos e das instituições de pesquisa. Nesse sentido, o paradigma social proposto por Capurro (2003) oferece uma base teórica potente para analisar os programas de pós-graduação como espaços de produção, organização e circulação do conhecimento científico. O autor argumenta que os processos informacionais são construções sociais, nas quais os documentos — como as teses e dissertações — funcionam como mediações simbólicas entre sujeitos, instituições e campos discursivos. Essa perspectiva é reforçada por Shera (1977), ao propor a epistemologia social como estudo dos meios pelos quais a sociedade produz e integra o conhecimento, e por Frohmann (1992), ao criticar a visão individualista do paradigma cognitivo e defender uma análise discursiva dos sistemas de informação. Birger Hjørland (2002), por sua vez, contribui com a noção de análise de domínio, segundo a qual os campos informacionais são constituídos pelas práticas e valores das comunidades científicas. A partir desse referencial, os programas de pós-graduação podem ser compreendidos não apenas como estruturas acadêmicas, mas como comunidades produtoras de sentidos, cujos sistemas de formação, avaliação e comunicação científica refletem disputas epistemológicas e interesses sociais diversos.

Conforme descrito por Saracevic (1996), a Ciência da Informação é uma ciência social, ou seja, abrange um campo de prática profissional e de investigação científica que lida com a comunicação eficaz de informações e objetos de informação. O aspecto social da área torna-se evidente quando são realizados os registros do conhecimento, que são utilizados na comunicação entre os seres humanos, tendo em vista atender as necessidades de um dado contexto social, organizacional e individual, buscando tornar possível o uso da informação (Bates; Maack, 2009). Para Saracevic (1996, p. 43), a Ciência da Informação tem natureza de uma ciência social, "uma vez que a ciência e a tecnologia são críticas para a sociedade", contexto de origem dessa disciplina. Para o autor, é "também crítico prover os meios para o fornecimento de informações relevantes para indivíduos, grupos e organizações envolvidas com a ciência e a tecnologia" (1996, p. 43). Citando Wersig e Neveling (1975), Saracevic (1996, p. 43) afirma que a Ciência da Informação "desenvolveu-se historicamente porque os problemas informacionais modificaram completamente sua relevância para a sociedade", sendo a responsabilidade social o fundamento da Ciência da Informação.

Borko (1968) já havia sinalizado a característica social da Ciência da Informação, ao descrever que ela se apresenta como ciência pura e aplicada. No primeiro caso, o conhecimento é produzido a partir de uma base intelectual, que estimula a reflexão, sem, necessariamente, possibilitar uma aplicação. Já no segundo caso, o autor citado declarou que, como ciência social aplicada, a Ciência da Informação desenvolve produtos e serviços direcionados a diferentes contextos organizacionais.

Para entender o aspecto de aplicação da Ciência da Informação, Borko (1968) descreveu que a área se destaca ao gerar produtos e serviços informacionais. Com o mesmo entendimento, Souza e Stumpf (2009) reforçaram a natureza aplicada da área, valorizando-a, também, por sua capacidade em propor métodos e técnicas que garantam a organização e o uso da informação pela sociedade.

Ao tratar de questões voltadas para a geração, comunicação e apropriação do conhecimento, [a Ciência da Informação] abrange um largo espectro de possíveis temáticas de interesse por estudiosos oriundos de diferentes áreas de formação acadêmica. Consequentemente, como campo de pesquisa e ensino, apresenta uma riqueza ímpar de possíveis metodologias de abordagem e propicia o desenvolvimento de diferentes processos, métodos e técnicas de coleta, tratamento e recuperação da informação (Souza; Stumpf, 2009, p. 42).

Considerando que se trata de uma ciência interdisciplinar na visão de alguns autores (Gomes, 2001; Saracevic, 1995; Silva *et al.*, 2006; Paim *et al.*, 2001), é possível que as descobertas oriundas do processo de pesquisa gerem produtos e serviços ou demais contribuições para diferentes profissões (Pinheiro, 2000). No caso da prática profissional dos bibliotecários, a Ciência da Informação compartilha o seu papel social e sua preocupação com os problemas da efetiva utilização dos registros gráficos, tal como se percebe na organização de acervos, sejam em bibliotecas, arquivos, museus ou outras unidades de informação (Queiroz; Vilan Filho; Moura, 2015). Quanto ao trabalho dos profissionais da comunicação, o propósito é cooperar para a transmissão da informação, tendo em vista os interesses comerciais e empíricos de pessoas e organizações (Queiroz; Vilan Filho; Moura, 2015).

De acordo com Pinheiro (2012), a Ciência da Informação é uma ciência aplicada em virtude de sua relação com outras áreas, estudando diferentes aspectos, como: contextos, áreas, setores e organismos. Isso quer dizer que a Ciência da Informação, por meio de uma abordagem teórica e prática, fomenta práticas profissionais, na sociedade, e as diferentes modalidades de informação, como informação científica, tecnológica, industrial ou artística, mesclam-se com a interdisciplinaridade propriamente dita (Pinheiro, 2012).

O estudo de Pinheiro (2012) comprovou essa aplicação da área, mediante o mapeamento dos PPGCI brasileiros, identificando as temáticas investigadas e as relações com

as diversas profissões existentes. De modo similar, Lança, Amaral e Gracioso (2018) também reforçaram a aplicabilidade da área, confirmando a forte aderência interdisciplinar e identificando a existência de pesquisas com diferentes abordagens, sejam elas de natureza teórica ou aplicada, além da constatação de que os PPGCI vêm crescendo, conforme mencionado na seção anterior.

De acordo com dados registrados na Plataforma Sucupira (Capes, 2022d), dentre os 27 PPG brasileiros na área da Ciência da Informação ativos, o Mestrado é o curso com maior quantidade, seguido pelo Doutorado e pelo Mestrado Profissional. Nota-se que os cursos predominantes são os de natureza acadêmica (Mestrado e Doutorado) e os cursos profissionais vêm sendo implementados nas instituições, embora o Doutorado Profissional ainda não tenha sido instituído na Ciência da Informação no país. A quantidade de cursos dos PPGCI brasileiros ativos pode ser visualizada no Quadro 8.

Quadro 8 – Número de cursos oferecidos nos 27 Programas em Ciência da Informação e os anos de implementação do primeiro curso

Curso	Quantidade	Ano de implementação do primeiro curso
Mestrado (ME):	19	1978
Doutorado (DO):	13	1978
Mestrado Profissional (MP):	8	2012
Doutorado Profissional (DP):	0	XXXX
Total geral:	40	

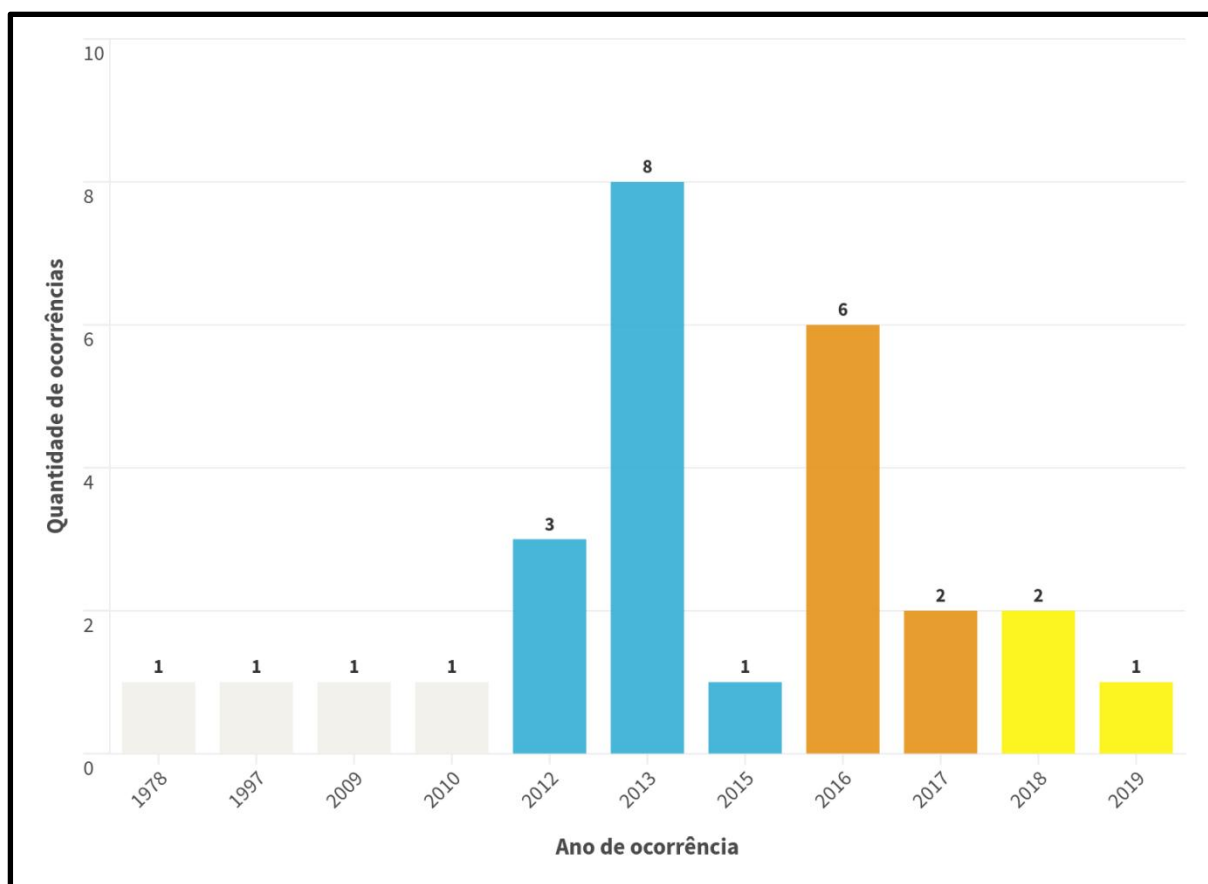
Fonte: Adaptado de Plataforma Sucupira (Capes, 2022d).

Pelas evidências dos Quadros 7 e 8, revela-se que a maioria dos Programas de pós-graduação (13 programas) oferecem cursos acadêmicos nas duas modalidades (Mestrado e Doutorado), ao passo que seis programas oferecem, apenas, o curso de Mestrado Acadêmico e oito programas somente o curso de Mestrado Profissional. Uma curiosidade é que, dentre o número de cursos oferecidos, apenas, uma modalidade de ensino, a modalidade Mestrado Profissional é maior do que a modalidade Mestrado Acadêmico.

Analisando os anos de implementação de cada programa, identifica-se o intervalo de crescimento. Assim, para os 27 programas ativos, o intervalo foi de 1978 a 2019, sendo que, a partir de 2009, houve criação de programas em quase todos os anos e, nos cinco últimos anos, a implementação foi consecutiva. Quanto aos cursos, a modalidade Mestrado Profissional surgiu no ano de 2008, com o Programa de Pós-Graduação em Gestão da Informação, da Universidade Estadual de Londrina (UEL), que foi oferecido somente até 2010 (Martins *et al.*,

2015). A data no Quadro 8 se refere à implementação do curso de Mestrado Profissional que ainda está vigente, que foi em 2012. É provável que os programas com, apenas, essas duas modalidades tendam a se expandir, por meio da realização de pesquisas mais profundas, o que possibilita a implementação da modalidade Doutorado. O Gráfico 2 apresenta o número de PPGCI, implementados ao longo dos anos, com identificação dos anos em que os cursos de uma modalidade foram implementados (Mestrado e/ou Mestrado Profissional).

Gráfico 2 – Número de Programas de Pós-Graduação – destaque aos anos com ocorrência do Mestrado e/ou Mestrado Profissional



Fonte: Adaptado de Plataforma Sucupira (Capes, 2022d).

LEGENDA: ■ Anos com ocorrência de Mestrados Profissionais
■ Implementação, apenas, dos Mestrados
■ Ocorrência do Mestrado Profissional e do Mestrado

Como mencionado, o Brasil possui oito PPG na área da Ciência da Informação que oferecem cursos na modalidade Mestrado Profissional. A fim de identificar o propósito de aplicação da pesquisa realizada nesses programas, apresentam-se os respectivos cursos, a saber:

1. **Mestrado Profissional em Biblioteconomia:** instituído no ano de 2012, na UNIRIO, esse curso objetiva possibilitar aos bibliotecários e profissionais da

informação a investigação relacionada a problemas oriundos do cotidiano de trabalho realizado nas bibliotecas, centros e unidades de informação, com base em questões de natureza teórica e empírica, sobretudo contemplando as técnicas, metodologias e práticas específicas do campo da Biblioteconomia (UNIRIO, 2022a);

2. **Mestrado Profissional em Gestão de Documentos e Arquivos:** esse curso também é ofertado na UNIRIO e tem seu escopo voltado à Arquivologia. Criado em 2012, o propósito é formar mestres preparados para atuar com a gestão de documentos e arquivos em serviços e instituições, sejam elas públicas ou privadas, atendendo as várias demandas sociais, profissionais e organizacionais do mercado de trabalho (UNIRIO, 2022b);
3. **Mestrado Profissional em Gestão da Informação:** oferecido na UDESC, o curso foi instituído no ano de 2013 e visa à formação especializada de profissionais comprometidos com a ética e a cidadania, tendo em vista resolver problemas associados à gestão da informação nas organizações, com base na inovação (UDESC, 2022);
4. **Mestrado Profissional em Gestão da Informação e do Conhecimento:** no ano de 2015, a UFRN implementou esse curso. Uma das intenções do curso é capacitar os profissionais no que tange à ampliação de conhecimentos técnico-científicos que garantam a solução de problemas e o desenvolvimento de produtos, processos ou serviços (UFRN, 2022);
5. **Mestrado Profissional em Memória e Acervos:** a FCRB, no Rio de Janeiro, criou esse curso no ano de 2016. O objetivo é formar gestores para atuação nos acervos de natureza arquivística, museológica ou bibliográfica, tendo em vista organizar a documentação produzida pelas instituições (FCRB, 2022);
6. **Mestrado Profissional em Gestão da Informação:** o curso é oferecido na USP, foi criado em 2016 e tem como proposta convergir teoria e prática, mediante o uso dos referenciais teóricos da área, na intenção de preparar os profissionais a desenvolver projetos de intervenção para o mundo do trabalho (USP, 2022);
7. **Mestrado Profissional em Biblioteconomia:** esse curso foi implementado no ano de 2016 e é oferecido na UFCA, cuja missão é aprofundar os conhecimentos científicos, na formação de pesquisadores e no envolvimento com a Biblioteconomia e valorizando a interdisciplinaridade, de modo a

fortalecer profissionais e pesquisadores da região (Região Metropolitana do Cariri, estado do Ceará) (UFCA, 2022);

8. **Mestrado Profissional em Gestão da Informação e do Conhecimento:** o curso passou a ser oferecido na UFS a partir de 2017, com a ideia de formar pesquisadores e profissionais preparados para atuar no campo da Ciência da Informação e áreas correlatas (UFS, 2022).

Pela análise ao objetivo/propósito de cada curso, evidencia-se o teor aplicado da pesquisa desenvolvida. Nota-se que, na maioria dos cursos, o propósito é a formação de profissionais para atuar no mercado, considerando temas que permeiam as práticas realizadas por profissionais que lidam com a informação, como: gestão da informação e do conhecimento, organização de acervos, centros de memória, gestão de documentos, dentre outras. Nesse sentido, os cursos, oferecidos em instituições brasileiras de prestígio, possibilitarão “[...] a qualificação de profissionais fomentando a produtividade/competitividade, via pesquisa, no campo da prática de mercado” (Carvalho, 2021, não paginado).

O percurso de nascimento e desenvolvimento dos PPG na área de Ciência da Informação e de seus respectivos cursos possibilita confirmar que a prática da pesquisa na Ciência da Informação é uma tendência relativamente nova, no contexto brasileiro. Mas, essa tendência é mundial, visto que a área se consolidou a partir de três importantes acontecimentos históricos: os avanços das tecnologias, das relações interdisciplinares e do nascimento da chamada sociedade da informação (Saracevic, 1996). De fato, essa ciência é nova porque “[...] suas dimensões humanas e sociais vão além das tecnologias, fornecendo uma visão mais ampla sobre a realidade baseada nos recursos tecnológicos, essencial para a compreensão e desenvolvimento da sociedade da informação” (Roza, 2018, p. 177).

Além disso, presume-se que o nascimento dos programas que oferecem cursos profissionais possa ser um indício que comprova a aplicabilidade social dessa área de conhecimento. Justifica-se essa afirmação a partir da essência de um PPG de natureza profissional, conforme previsto no V PNPG, de que as descobertas científicas precisam estar alinhadas com a aplicação de conhecimento. Isso quer dizer que “[...] o conhecimento passa cada vez mais a ser visto como fruto de um processo dinâmico heterogêneo voltado para o contexto de aplicação” (Barbosa, 2020, p. 5).

O curso de Mestrado Profissional na área de Ciência da Informação surge como uma necessidade do mercado de trabalho. Martins *et al.* (2015, p. 411) relataram que “[...] a crescente procura e necessidade de aperfeiçoamento dos profissionais graduados fez com que

buscassem a obtenção de novos conhecimentos para qualificar a sua atuação profissional [...]”. As autoras também afirmaram que a oferta do curso de Mestrado Profissional em Ciência da Informação, no Brasil, é recente e que esse curso não visa à formação de pesquisador, como o Mestrado e Doutorado acadêmicos, mas à formação para o mundo profissional, a partir da imersão na pesquisa aplicada, proporcionando uma formação que torne o profissional criativo e apto a melhorar o contexto onde atua.

Portanto, a criação dos cursos profissionais na Pós-Graduação em Ciência da Informação pode ser um indicativo de que a área é cada vez mais requisitada no mundo do trabalho, revelando a sua aplicação na sociedade. Como sinalizou Le Coadic (2004, p. 21), a Ciência da Informação procura “[...] esclarecer um problema social concreto, o da informação, sendo voltada para o ser social que procura informação [...]”. Assim, a pesquisa em Ciência da Informação é orientada a responder uma necessidade social, contemplando o mundo do trabalho, das relações sociais e culturais (Le Coadic, 2004).

Em resumo, confirma-se que a área da Ciência da Informação, no Brasil, está em desenvolvimento, tanto pela existência dos PPG, sejam eles acadêmicos ou profissionais, como por meio dos periódicos que possibilitam a transmissão dos resultados das pesquisas, além da vigência das associações profissionais e dos eventos científicos. De qualquer modo, a qualidade do que é produzido pela área precisa ser mantida e tal processo pode ser realizado por meio de técnicas específicas de avaliação e de mensuração, como os estudos bibliométricos (Silva, 2019; Oliveira, 2021). Os estudos bibliométricos estão inseridos no campo denominado por Saracevic (2000) de Métricas da Informação, campo esse, também, de responsabilidade da Ciência da Informação. As concepções conceituais sobre esses estudos e os principais aspectos ou elementos estudados dentro da Ciência da Informação são descritas no subcapítulo seguinte.

2.4 Estudos bibliométricos na área de Ciência da Informação

Ao contextualizar a Ciência da Informação com descrições acerca dos conceitos, características e desenvolvimento dessa área, Saracevic (2009) apontou os estudos bibliométricos, revelando que se trata de um conjunto de técnicas que exploram as propriedades estatísticas de determinados elementos. Menciona o autor que a finalidade desses estudos é conhecer as relações e princípios de uma variedade de entidades dentro de um domínio de conhecimento, revelando o desenvolvimento e o comportamento dessas entidades.

Originados após a década de 1950, acompanhando a evolução da área da Ciência da Informação, os estudos bibliométricos, ou Bibliometria, foram caracterizados como “[...] o estudo quantitativo das propriedades da literatura, ou mais especificamente, de documentos e processos relacionados a documentos [...]” (Bates; Maack, 2009, p. 2575, tradução nossa). De modo parecido, Fonseca (1986) destacou que a Bibliometria, a princípio, foi associada à contagem de livros e outros materiais de biblioteca, mas que extrapolou essa concepção, ao explorar os relacionamentos e variáveis oriundos das técnicas estatísticas.

Çelik (2020) defende que a Bibliometria é um método de quantificação, que analisa a literatura de modo minucioso, possibilitando diferentes combinações entre variáveis, por meio de técnicas matemáticas e estatísticas. O autor valoriza esse método, pois ele tende a mostrar informações que avaliam o desempenho de uma variável ou elemento, sendo, portanto, um indicador que pode qualificar o que está sendo mensurado. Portanto,

As publicações e pesquisadores mais efetivos, revistas populares sobre assuntos específicos, e países e instituições proeminentes podem ser determinados com a análise bibliométrica. Portanto, pesquisadores que desejam acompanhar as mudanças na literatura acadêmica de um assunto específico e se informar sobre a literatura geral podem usar este método de análise. A análise bibliométrica é um método proeminente usado no processo de tomada de decisão científica, e também é amplamente utilizada para avaliar o desempenho de revistas, países e instituições (Çelik, 2020, p. 3632, tradução nossa).

Percebe-se que a Bibliometria está associada à quantificação da produtividade científica, *a priori*, podendo gerar resultados muito mais abrangentes, relacionados a aspectos não quantificáveis que evidenciem a qualidade, a atividade e o impacto científico, além das associações temáticas (Beira *et al.*, 2020). Assim, os estudos bibliométricos são fundamentais por levantarem os indicadores de avaliação da produção científica, manifestando-se, portanto, de grande interesse para o mapeamento da produção científica brasileira. Os resultados obtidos na análise bibliométrica dependem da qualidade dos dados coletados em sistemas de recuperação da informação e, para que não haja inconsistências nos dados coletados, “[...] é preciso que as informações tenham passado por processo consistente de organização” (Silva *et al.*, 2013, p. 259).

Embora tenha se originado na Ciência da Informação e se constitui um forte campo de pesquisa nessa área, a Bibliometria vem sendo utilizada em domínios do conhecimento, tais como na Comunicação (Vanz, 2002) e na Administração (Caldas; Tinoco, 2004). Em específico, na área das Ciências Sociais Aplicadas, os estudos bibliométricos analisam a produção de artigos em um domínio, as relações com outros domínios, as temáticas abordadas, além de revelar as comunidades ou personalidades envolvidas nessas produções (Chueke;

Amatucci, 2015). Com isso, tornam-se esclarecidos os pontos de aproximação entre as áreas, os relacionamentos firmados e as motivações que levam a essas relações. Os resultados dessas medições “[...] são tangibilizados por meio da criação de indicadores que buscam sumarizar as instituições e os autores mais prolíferos, os acadêmicos mais citados e as redes de coautorias” (Chueke; Amatucci, 2015, p. 2).

A aplicação do método em pesquisas da Ciência da Informação tem revelado o quanto a área tem crescido e as relações que acontecem com outras áreas do conhecimento vão se expandindo. Justifica-se essa afirmação, tendo em vista que a Bibliometria é usada para medir não apenas a produção de pesquisa ou produtividade de autores individuais, organizações e países, mas também a extensão da colaboração entre eles. Pesquisa e desenvolvimento são as atividades centrais em qualquer campo científico. Em resposta a essa necessidade, os pesquisadores de distintas áreas desenvolveram pesquisas sobre todos os aspectos que envolvem as profissões dessas áreas (Siddique; Rehman; Khan, 2021).

A avaliação e a avaliação desempenham um papel fundamental na identificação dos pontos fortes e fracos de qualquer campo, o que ajuda na tomada de decisões sólidas e no desenvolvimento de políticas. Ao longo dos anos, os pesquisadores adotaram técnicas quantitativas como a bibliometria para investigar a produtividade e o impacto de diferentes aspectos relacionados à Biblioteconomia e Ciência da Informação. (Siddique; Rehman; Khan, 2021, p. 90, tradução nossa).

Com o desenvolvimento tecnológico e o crescimento exponencial do número de documentos produzidos, a Ciência da Informação e a Biblioteconomia uniram esforços para desenvolver métodos e técnicas capazes de mensurar e avaliar a produção documental contribuindo, assim, para demonstrar o crescimento da ciência e estabelecer indicadores de qualidade (Alvarez; Caregnato, 2017). Esses autores também pontuam que a criação de bases de dados como, por exemplo, a Web of Science, “[...] foi fundamental para a realização de estudos e a obtenção de indicadores bibliométricos. A formulação de leis e teorias ajudou a compreender o comportamento e a composição da literatura publicada nos periódicos das diferentes áreas” (Alvarez; Caregnato, 2017, p. 23).

Assim, confirma-se que a Ciência da Informação, por meio da Bibliometria, exerce valiosa contribuição para a medição do conhecimento científico, por conseguinte, facilita a avaliação desse conhecimento (Alvarez; Caregnato, 2017). Silva (2019) corrobora essa afirmação e reforça que o principal contributo está na capacidade de medir a produtividade das áreas do conhecimento, representadas por instituições, por pesquisadores, por periódicos, por PPG, dentre outros elementos.

Ao analisar a produtividade dos PPG em Ciência da Informação, por meio de estudo bibliométrico, Silva (2019) concluiu que a produtividade deve estar associada à qualidade, “[...] porque o reflexo que o trabalho produz na ciência permite sua interdiscursividade além das cercas de produto científico; seus resultados resultam em benesses para sociedade como um todo [...]”. Também constatou que o estímulo à produtividade decorre da necessidade de se cuidar para que os pesquisadores não adoeçam de “normose acadêmica”, ou seja, que os pesquisadores não se acomodem em suas práticas cotidianas, fato esse que pode prejudicar a criatividade e a livre iniciativa para o desenvolvimento de novos conhecimentos (Silva, 2019, p. 168).

De fato, a Bibliometria como técnica de investigação para as pesquisas da Ciência da Informação irá mostrar a gênese da área, o seu percurso e as relações estabelecidas (Oliveira, 2021). Desse modo, segundo o mesmo autor, será possível identificar o progresso científico e como ocorreu a produção de novos conhecimentos científicos para a sociedade, além de indicar a formação das novas gerações de pesquisadores.

Assim como Oliveira (2021), Coelho (2021) também defende a abrangência de uso dos estudos bibliométricos para identificação da área da Ciência da Informação, enfatizando que diversas variáveis podem ser medidas, incluindo-se as redes de relacionamento entre os pesquisadores para desenvolvimento das pesquisas. Na formação dessas redes, destacam-se, dentre outras possibilidades, a identificação de temáticas de estudo, as tecnologias e as técnicas utilizadas, como também a visualização de dados, conforme será detalhado nas subseções seguintes.

Diante da importância atribuída à bibliometria no campo da Ciência da Informação, como técnica de análise da produção científica e instrumento de avaliação de desempenho, torna-se necessário aprofundar sua fundamentação no contexto desta pesquisa. A seguir, está apresentada uma breve discussão teórico-metodológica que integra os fundamentos epistemológicos que orientam a bibliometria e a curadoria de dados como práticas informacionais. Essa base permite compreender, de forma articulada, tanto as decisões metodológicas adotadas quanto os procedimentos técnicos apresentados nas subseções seguintes, que tratam das temáticas abordadas, das leis e técnicas bibliométricas, dos modos de visualização e das redes de colaboração científica.

2.4.1 Fundamentos epistemológicos: bibliometria e curadoria de dados como práticas informacionais

A fundamentação epistemológica adotada nesta pesquisa se apoia em autores contemporâneos da Ciência da Informação, especialmente a partir dos anos 2000, cujas contribuições permanecem centrais nos debates atuais da área. Trata-se de pensadores que, embora tenham produzido parte significativa de suas obras entre o final da década de 1990 e os anos 2010, continuam sendo debatidos por sua capacidade de iluminar aspectos estruturais das práticas informacionais. Nesse conjunto destacam-se Frohmann (2004), Capurro (2007), Cyr e Wei Choo (2010), Buckland (2012), Borgman (2015) e Hjørland (2013), cujas abordagens compreendem a informação como construção social, os dados como artefatos interpretativos e os documentos como mediadores entre sujeitos e saberes. Ao mobilizar essas contribuições, busca-se reforçar a compreensão de que tanto a bibliometria quanto a curadoria de dados não são apenas operações técnicas, mas práticas informacionais situadas, que refletem valores, escolhas epistemológicas e condições institucionais.

A abordagem adotada nesta pesquisa, que combina técnicas bibliométricas com práticas de curadoria de dados, encontra respaldo em debates epistemológicos contemporâneos da Ciência da Informação. Tais debates problematizam a compreensão da informação como uma entidade abstrata e autônoma, deslocando o foco para os documentos, as práticas e os contextos de produção e uso da informação.

Frohmann (2004), por sua vez, critica a concepção abstrata de informação dominante nos estudos sobre comunicação científica, especialmente aqueles que tratam documentos como meros recipientes de conteúdo informacional. Para o autor, essa visão desconsidera as práticas concretas e materiais envolvidas na produção, circulação e uso dos documentos. Ele argumenta que "a vida dos documentos, assim como a vida das palavras, está nos usos que deles são feitos", defendendo uma abordagem documentalista que desloca o foco dos significados abstratos para as práticas e os contextos em que os documentos ganham sentido (Frohmann, 2004, p. 16).

Essa concepção também dialoga com o paradigma social proposto por Capurro (2007), segundo o qual os dados informacionais devem ser compreendidos como construções socialmente situadas, inseparáveis do contexto cultural e das práticas discursivas em que são produzidos. O autor destaca que só faz sentido falar de informação quando esta é interpretada como nova e relevante para uma comunidade ou grupo social específico.

De modo convergente, Cyr e Wei Choo (2010) aprofundam a análise das dinâmicas de compartilhamento do conhecimento, já antes discutidas por Wei Choo, destacando que tais práticas não se reduzem a mecanismos técnicos ou automatizados, mas envolvem julgamentos sociais mediados por percepções de valor, reciprocidade e normas culturais. O modelo proposto pelos autores identifica o compartilhamento como uma forma de ação social intencional, sensível à orientação para o outro, ao tipo de vínculo existente entre os envolvidos e ao equilíbrio percebido entre custos e benefícios. Essa perspectiva fortalece a ideia de que a curadoria de dados, tal como concebida nesta pesquisa, é uma prática relacional situada, que reflete valores, interpretações e formas específicas de engajamento epistêmico e social. Assim, tanto Capurro quanto Cyr e Wei Choo (2010) reforçam que as práticas informacionais — como a curadoria e a bibliometria — não são neutras, mas constituem modos de construção e gestão do conhecimento em contextos historicamente determinados.

Por sua vez, Buckland (2012) destaca que a informação pode ser compreendida em três sentidos: informação como processo, como conhecimento e como coisa. Esta última, também chamada de informação-como-coisa, corresponde a objetos físicos (como textos, dados, imagens, registros) que são percebidos como portadores de significado. Ao afirmar que "a informação não é uma coisa, mas um processo de engajamento cultural", Buckland propõe uma compreensão da Ciência da Informação como uma "ciência do artificial", voltada às condições culturais e institucionais de produção e uso dos documentos (Buckland, 2012, p. 7).

Diferentemente da aproximação à filosofia da informação e da tradição documentalista de Buckland (2012), Hjørland (2013) traz um enfoque vinculado à perspectiva sociocognitiva e à teoria dos domínios. O autor sustenta que a Ciência da Informação deve se basear na análise dos domínios do conhecimento e das práticas que os constituem. Hjørland (2013) enfatiza que o significado da informação depende dos contextos sociais, históricos e epistemológicos em que é produzida e utilizada, sendo, portanto, inseparável das condições de produção do conhecimento. Ao abordar o conceito de domínio, o autor propõe que a organização e representação da informação devem estar fundamentadas no entendimento dos valores, finalidades e formas de conhecimento predominantes em cada campo (Hjørland, 2013).

Complementarmente, Borgman (2015) defende que os dados não são objetos naturais, mas artefatos construídos social e tecnicamente. Em sua análise sobre a infraestrutura do conhecimento, a autora afirma que "os dados devem ser compreendidos como entidades situadas, cujo valor, significado e uso variam conforme o campo, o tempo e o local em que são gerados" (Borgman, 2015, p. 33). Assim como Hjørland ressalta a importância dos

contextos epistemológicos e Buckland enfatiza o engajamento cultural, Borgman reforça que a materialidade dos dados e os processos de curadoria são moldados por práticas institucionais e valores científicos. Para ela, a curadoria de dados é uma prática central da ciência contemporânea, especialmente quando se busca promover transparência, reprodutibilidade e reuso, valores fundamentais da ciência aberta. Sua abordagem enfatiza que compreender a informação — e, mais especificamente, os dados — exige atenção às infraestruturas sociotécnicas que viabilizam sua circulação, acesso e interpretação.

Essas perspectivas sustentam a compreensão de que tanto a bibliometria quanto a curadoria de dados devem ser concebidas como práticas informacionais inseridas em contextos epistemológicos, sociais e institucionais. Na presente pesquisa, a curadoria de dados não é entendida como uma etapa meramente técnica da coleta, mas como uma atividade que envolve escolhas interpretativas e críticas sobre o que considerar como dado e como representá-lo. Ao adotar esse enquadramento, busca-se alinhar as opções metodológicas às discussões contemporâneas da Ciência da Informação, reconhecendo os dados e os documentos como produtos culturais mediados por condições de produção, circulação e uso.

Complementando esse enquadramento contemporâneo, é igualmente importante reconhecer que tais práticas informacionais se inscrevem em uma tradição teórica mais ampla da Ciência da Informação. Desde a definição inaugural proposta por Borko (1968), que concebe a área como um campo interdisciplinar voltado à coleta, organização, armazenamento, recuperação e uso da informação, observa-se a constituição de uma base conceitual consistente sobre o papel da informação na produção científica. Autores como Shera (1977) também contribuíram com o conceito de epistemologia social, reforçando a ideia de que os documentos — como as teses e dissertações aqui analisadas — são produtos de comunidades de saber e expressam valores, conflitos e consensos coletivos. Ao reconhecer essa herança, a presente pesquisa busca não apenas dialogar com os debates mais recentes sobre curadoria de dados e bibliometria, mas também afirmar sua vinculação com o núcleo teórico fundante da área.

Com base nas perspectivas epistemológicas que sustentam a compreensão da informação como construção social e dos dados como artefatos interpretativos, a presente pesquisa avança agora para o referencial teórico-metodológico específico dos estudos bibliométricos.

2.4.2 As temáticas abordadas

Mapear as temáticas investigadas em uma área de conhecimento por meio dos estudos bibliométricos é uma estratégia de conhecer melhor o campo e seus limites de interação com outros campos e disciplinas. Na Ciência da Informação, esse mapeamento é necessário, pois possibilita conhecer as delimitações do campo, a abrangência das relações, com a finalidade de apresentar um panorama temático da área, com revelações sobre temas recentes e futuros contemplados nas pesquisas, sendo a pesquisa, pois, dispersa (Deng; Xia, 2020).

Estudos realizados em países externos ao Brasil têm adotado a análise dos temas publicados em periódicos, visto que boa parte da produção dos PPG e das experiências profissionais está registrada nesses canais de comunicação. Na Ásia, por exemplo, o *Annals of Library and Information Studies*, periódico de abrangência em todo o continente e sediado na Índia, foi analisado, cuja fonte dos dados utilizada foi a base de dados multidisciplinar Scopus. Diversas temáticas inerentes ao campo da Ciência da Informação foram identificadas, com destaque pesquisas sobre as Métricas da Informação, Bibliometria, redes de colaboração e redes sociais (Prieto-Gutiérrez; Segado-Boj, 2019). Foi concluído, também, que essas temáticas são uma tendência e conseguem representar as instâncias globais sobre o campo temático da Ciência da Informação.

Outra estratégia para mapear as temáticas pesquisadas na Ciência da Informação é recorrer à coleta em diferentes fontes de dados, tais como: base de dados de abrangência mundial, como a Web of Science e Scopus, e bases de escopo especializado, como Library and Information Science Abstracts, Library and Information Science and Technology Abstract e Educational Resources Information Center (Pinto *et al.*, 2019). Por meio de técnicas estatísticas, foi possível, com o citado estudo, identificar as temáticas mais recorrentes da área, a partir de um período estudado, e delimitar o escopo temático de boa parte da Ciência da Informação mundial. Esse estudo também adotou a delimitação temática, ou seja, investigou a evolução de um tema da Ciência da Informação, tal como letramento informacional, revelando que esse tema é representativo nas pesquisas realizadas no contexto dessa área de conhecimento em diversos países.

Também em nível internacional, muitos estudos têm analisado a produção temática publicada em periódicos especializados ou anais de eventos da área da Ciência da Informação, tal como realizado no International Journal of Information Management, em decorrência dos 40 anos de implementação do periódico (Donthu *et al.*, 2021). Os resultados indicaram que os temas mais dominantes foram: gestão de sistemas de informação, tecnologias da informação e

sua aplicação nos negócios, aceitação de tecnologia entre os consumidores, uso de sistemas de informação para tomada de decisão, perspectivas sociais sobre gestão do conhecimento e pesquisa da informação na perspectiva das Ciências Sociais. Ademais, ficou evidenciado que a maioria dos estudos é produzida no contexto dos Estados Unidos e da Europa (Donthu *et al.*, 2021).

Assim como no exterior, no Brasil, muitos estudos são desenvolvidos a partir da coleta de dados em bases de dados e em revistas especializadas. O estudo de Bufrem *et al.* (2007) utilizou de coleta quantitativa na BRAPCI, por meio da análise das temáticas publicadas nos periódicos indexados nessa base. Para esses autores, explorar o conteúdo dos periódicos tem sido uma modalidade de estudo com presença reiterada na literatura voltada à produção de conhecimento. Pelos resultados do estudo, concluiu-se que as temáticas são variadas e podem ser agrupadas em três grandes categorias: a produção científica propriamente dita, em suas manifestações; a análise temática dessa literatura; e as suas formas de expressão, tais como, periódicos, teses, dissertações e literatura cinzenta⁶.

Além das temáticas da área de Ciência da Informação serem identificadas, nos estudos bibliométricos, mediante bases de abrangência nacional, muitas pesquisas coletam dados em bases de dados de referência internacional e com escopo multidisciplinar, como as bases Web of Science e Scopus, conforme procedimentos adotados por Silva (2021). Foi concluído, nessa investigação, que os temas mais recorrentes são: Ciência da Informação, competência informacional e informação, embora a produção se caracterize como diversificada, com presença de diferentes periódicos, autores e instituições e há predomínio da prática da colaboração entre autores no desenvolvimento das pesquisas.

Assim como o estudo de Donthu *et al.* (2021) fez uma análise bibliométrica para mapear os temas publicados nos artigos de um periódico em específico, no Brasil, alguns outros estudos adotam essa mesma metodologia. Maia *et al.* (2020), por exemplo, levantaram dados da Revista Analisando em Ciência da Informação (RACIn), cujas temáticas identificadas representaram o que consideram a característica interdisciplinar da Ciência da Informação, contemplando assuntos de áreas correlatas, tais como: Administração, Arquivologia, Biblioteconomia, Educação e Informática. Também foi evidenciado que as análises bibliométricas em temáticas da Ciência da Informação permitem a elucidação do

⁶ O termo “literatura cinzenta” ou *grey literature* foi cunhado em 1978, no Seminário de York, no Reino Unido, e é definida como “publicações não convencionais e não comerciais, semi publicadas, difíceis de encontrar em canais tradicionais de distribuição, com controle bibliográfico ineficaz (não recebem numeração internacional e não são objeto de depósito legal em muitos países), sendo frequentemente não incluídas em bibliografias e catálogos” (Botelho; De Oliveira, 2015, p. 511).

volume da informação publicizada “[...] e suas tênues conexões e contribuições com as diversas áreas dos saberes, permitindo-se ao diálogo e às possibilidades epistemológicas de forma aberta” (Maia *et al.*, 2020, p. 45).

Além da análise ao conteúdo dos artigos publicados em periódicos, as abordagens bibliométricas podem contemplar os artigos publicados em anais de eventos. Nesse sentido, Mata (2021) analisou como amostra os artigos do Enancib e também delimitou uma temática específica no campo da Ciência da Informação: a competência em informação. Mediante a aplicação de um estudo exploratório e bibliográfico, com abordagem quantitativa e qualitativa, a conclusão foi que as pesquisas sobre a temática têm sido ampliadas ao longo dos anos no Enancib, cujo termo mais utilizado foi competência em informação. A autora também destaca que os desdobramentos de pesquisa precisam ser ampliados, tendo em vista que as temáticas da Ciência da Informação adentram outros campos de investigação.

Por fim, mas não menos importante, o mapeamento temático da Ciência da Informação no Brasil pode ocorrer, também, com grande frequência, por meio da análise às palavras-chave das teses e dissertações, que se referem aos temas representativos do texto original, em geral, extraídos do resumo e considerados “os mais expressivos a nível semântico” (Simões, 2015, p. 133). Com isso é possível correlacionar as temáticas às linhas de pesquisas dos PPG, e, por conseguinte, tem-se o panorama de pesquisa dos programas. Ao aplicar essa abordagem analítica em teses e dissertações de um PPG, Souza (2021) concluiu que é importante compreender as temáticas que envolvem a Ciência da Informação, pois possibilitam identificar as relações e interdisciplinaridades acerca das temáticas pesquisadas. Nesse estudo, também foi constatado como temáticas mais recorrentes: gestão de documentos, comunicação científica, Arquivologia e biblioteca universitária, o que evidenciou boa parte dos assuntos que compõem a Ciência da Informação brasileira.

Para Souza (2021), a realização das análises bibliométricas na Ciência da Informação requer um conjunto de procedimentos, representado por leis, técnicas e tecnologias específicas que consolidam tais processos como metodológicos. É essa diversidade de procedimentos que tornam os estudos bibliométricos, segundo Beira *et al.* (2020), cada vez mais utilizados, com resultados que revelam contribuições quantitativas e qualitativas, visando à qualidade da produção científica. As leis, técnicas e tecnologias mais recorrentes nos estudos bibliométricos são apresentadas na próxima subseção.

2.4.3 Leis, técnicas e tecnologias

Dentre os procedimentos adotados na Bibliometria, destacam-se estudos desenvolvidos sob o enfoque das leis bibliométricas. Essas leis apresentam-se como princípios de comportamento, no sentido de que a produção científica pode ser identificada e avaliada, haja vista atender a padrões de desempenho estabelecidos pelas instituições responsáveis (Machado Júnior *et al.*, 2016). Com efeito, as leis foram se desenvolvendo com uso de recursos de tecnologia de ponta, sendo propostos modelos de comportamento que se estabeleceram em padrões de análise de dados. Assim, as principais leis estabelecidas com esse desenvolvimento da Bibliometria foram: Lei de Lotka, Lei de Brandford e Lei de Zipf (Machado Júnior *et al.*, 2016).

No entendimento de Kurtuluş e Tatar (2021), as leis bibliométricas também denominadas de modelos bibliométricos são as regras estabelecidas com base em vários cálculos e distribuições estatísticas, cuja finalidade principal é avaliar os processos de pesquisa. Nesse contexto, segundo esses autores, podem ser analisadas, por meio das leis: a distribuição dos artigos, a média de pontuação de citação, a lista dos periódicos mais produtivos, a lista dos autores mais produtivos e citados, as citações dos autores, a produtividade científica dos países e dos autores e as redes de colaboração, variáveis essas que podem ser obtidas por meio de mineração de textos. Portanto, “[...] essas leis são geralmente funções matemáticas e estatísticas simples que analisam a correlação de uma variável com outra variável [...]” (Kurtuluş; Tatar, 2021, p. 195, tradução nossa).

As leis bibliométricas possibilitam a análise quantitativa de partes da literatura acadêmica, com o propósito de medir o desempenho de alguns elementos que formam a literatura, como autores, periódicos e o conteúdo textual representado em palavras e expressões (Sharma; Chakravarty, 2018). Proferem os autores que essas leis são muito parecidas, com diferenças marcantes no elemento analisado e no tipo de dados. Assim, a lei da produtividade científica de Lotka mensura os autores das publicações de uma determinada disciplina, a lei de Bradford explora a dispersão dos artigos e a lei de Zipf analisa as palavras (*ranking* da frequência da palavra).

Considerando o que difere uma lei de outra, Sharma e Chakravarty (2018) mencionam que a lei de Lotka trata das publicações dos autores e do número de artigos publicados; Bradford observou a divulgação de artigos sobre assuntos específicos em vários periódicos; e Zipf contou as frequências das palavras. Para melhor detalhamento sobre conceitos, características e aplicações das três leis, apresenta-se o Quadro 9.

Quadro 9 – Conceitos e aplicações das leis da Bibliometria

Lei	Breve descrição	Principal aplicação
Lei de Lotka	A Lei de Lotka ou Lei do Quadrado Inverso propõe que um número restrito de pesquisadores produz muito em determinada área de conhecimento, enquanto um grande volume de pesquisadores produz pouco. Lotka estudou os autores presentes no Chemical Abstracts, entre 1909 e 1916, e identificou que grande parte da produção científica é produzida por poucos autores. A produção desse número reduzido de autores se iguala em quantidade ao desempenho de muitos autores que possuem baixo volume de publicação.	A representação desse princípio pode ser expressa matematicamente como o número de autores que publica n artigos (n é igual à quantidade de artigos) que é igual a $1/n^2$ dos autores que publicam somente um artigo. A lei pode ser aplicada na identificação da produtividade autoral, das colaborações entre autores, suas citações, dentre outras possibilidades.
Lei de Bradford	Incide sobre o conjunto de periódicos e surgiu de pesquisas médicas conduzidas por Hill Bradford e outros médicos do Conselho de Pesquisas Médicas Americano. A proposição do estudo era a de identificar a extensão de publicação de artigos científicos de um assunto específico, em revistas especializadas daquele tema. Os dados coletados mostraram a existência de um pequeno núcleo de periódicos que aborda o assunto de maneira mais extensiva, e uma vasta região periférica dividida em zonas. Nessas zonas, observa-se o aumento do número de periódicos que reduzem a produtividade de publicação de artigos do respectivo assunto.	Estima o grau de relevância de periódicos que atuam em áreas do conhecimento específicas. Periódicos com maior publicação de artigos sobre determinado assunto tendem a estabelecer um núcleo supostamente de qualidade superior e maior relevância nesta área do conhecimento. Essa lei se aplica à medição da produtividade dos periódicos e as possíveis variáveis que possam ser mapeadas envolvendo a produção periódica.
Lei de Zipf	Consiste em medir a frequência do aparecimento das palavras em vários textos, gerando uma lista ordenada de termos de uma determinada disciplina ou assunto.	Relaciona-se diretamente com a representação da informação, isto é, a indexação temática e automática. Tal lei pode ser aplicada na elaboração de resumos, na produção de conteúdo, dentre outros casos.

Fonte: Adaptado de Guedes e Borschiver (2005) e Machado Junior *et al.* (2016).

O Quadro 9 revela as diferenças entre cada lei, confirmando que cada uma realiza um tipo específico de mensuração, como descrito no estudo de Sharma e Chakravarty (2018). A especialidade de cada lei promoveu o seu uso em estudos bibliométricos de diferentes áreas do conhecimento, com destaque os estudos da própria Ciência da Informação. Isso porque é fornecido tratamento matemático e estatístico à informação registrada, o que contribui para sustentar a base teórica dessa área, conforme descrito na clássica citação de Price (1965, p. 510, tradução nossa), ao enfatizar que, pelas leis, “[...] torna-se possível alcançar resultados interessantes se aceitarmos que há uma relação entre a ciência e sua respectiva literatura, isto é, que há uma relação entre conhecimento e informação registrada.”.

Portanto, na área da Ciência da Informação, as principais leis da Bibliometria vêm sendo bastante utilizadas. A lei de Lotka foi utilizada no estudo de Sharma e Chakravarty (2018) para mapear a produtividade dos professores das faculdades de Biblioteconomia do norte da Índia, sendo analisada uma amostra composta por 1.292 produções (artigos, capítulos de livros, anais de evento e livros) de 30 professores. A técnica bibliométrica utilizada foi a fórmula matemática $y = C \times x^{-n}$, onde x é o número de publicações, n é um expoente constante para um dado conjunto de dados, y é a porcentagem esperada de autores com frequência x de publicações e C é uma constante. Pelos resultados alcançados com a aplicação da lei de Lotka, foi possível apresentar o *ranking* dos autores mais produtivos por instituição, concluindo que essa lei pode ser utilizada para comparar a produção dos autores das instituições acadêmicas.

No Brasil, a lei de Lotka foi testada na dissertação de Cunha (2021, p. 42), ao partir do princípio que tal lei “[...] analisa a produtividade de cientistas e indica que grande porção da literatura é produzida por um pequeno percentual de autores, quando a maioria dos autores é responsável por poucas publicações [...]”. Com o propósito de verificar o termo cultura no contexto da literatura científica periódica em Ciência da Informação, produzida pelos bolsistas de produtividade do CNPq, Cunha (2021) utilizou como técnicas de coleta e análise de dados, planilhas eletrônicas do *Microsoft Excel* e os *softwares Ucinet* e *Netdraw* para a geração de matrizes e redes, facilitando, pois, a visualização dos dados. Com esses procedimentos, a partir da análise aos autores, verificou-se que a Lei de Lotka não incidiu sobre o conjunto de dados, “[...] visto que Lotka indica em sua lei que, a proporção de autores responsáveis por apenas uma publicação deve corresponder a aproximadamente 60% da comunidade analisada [...]”, sendo que essa proporção foi bem menor na amostra analisada (Cunha, 2021, p. 58).

A lei de Bradford também foi aplicada em estudos desenvolvidos no Brasil, tal como no mapeamento da produção científica sobre o termo Governo Aberto na base Web of Science durante o período de 2010 a 2016, com a finalidade de identificar a produção veiculada nos periódicos da área de Ciências Sociais Aplicadas, incluindo a Ciência da Informação (Sembay *et al.*, 2020). Tecnicamente, foi aplicada a lei bibliométrica de Bradford e utilizou-se como ferramenta uma planilha eletrônica para analisar os dados obtidos na mineração, bem como tabulá-los e tratá-los. A partir dos resultados, os autores concluíram que, no decorrer do tempo, “[...] as publicações apresentam uma aparência não linear, registrando um aumento significativo da produção em relação ao contexto internacional [...]”. Ao contextualizar com a lei de Bradford, confirmou-se que “[...] o termo Governo Aberto

apresenta qualidade nas publicações científicas em que é difundido, mas que ainda está em fase de luta por seu espaço nas publicações científicas dentro de uma ciência atual em transformação” (Sembay *et al.*, 2020, p. 1).

Também no Brasil, a lei de Zipf aparece em alguns estudos, tal como em Ivo *et al.* (2021), cujo objetivo foi identificar a produção intelectual de dois pesquisadores registrados na Plataforma *Lattes*, autodeclarados como estudiosos do tema política de informação. O *framework LattesDataXplorer* (ferramenta composta por um conjunto de técnicas e métodos para coletar, selecionar, processar e analisar dados contidos em determinado currículo armazenado na Plataforma *Lattes*) e o *software Publish or Perish* (que recupera e analisa citações acadêmicas) foram utilizados como ferramentas tecnológicas para extração e análise de dados. Ao aplicar a lei de Zipf, no que tange à terminologia, foi constatado que 76 palavras compõem a primeira esfera da lei de Zipf, sendo a palavra Ciência da Informação representada $n=218$ vezes. Assim, foi concluído que

[...] o tema Política de Informação é o foco da Ciência da Informação evidenciado pela concentração e elite de dois pesquisadores, bem como a predominância de dois canais de comunicação e divulgação, não se destacando dentro das bases de dados, podendo haver limitações de linguagem e baixa capilaridade dos periódicos multidisciplinares (Ivo *et al.*, 2021, p. 83).

Percebe-se que as leis da Bibliometria apresentam diversas técnicas que possibilitam extrair os dados das fontes e analisá-los, cujos procedimentos são complementados com o uso de ferramentas tecnológicas desenvolvidas para auxiliar as análises bibliométricas, tais como os programas de extração dos dados das fontes consultadas (Souza, 2021). “Por meio de seus instrumentos e métodos, a Bibliometria possibilita reconhecer o comportamento de determinado campo científico [...]”, cujas técnicas “[...] podem ser aplicadas em documentos distintos, sendo possível a análise de diversos dados que podem gerar diferentes indicadores [...]”, promovendo, assim, a avaliação da produção científica (Souza, 2021, p. 27).

Durante os processos de extração e análise, os dados são expostos, permitindo a sua visualização, por conseguinte, torna-se possível identificar, reconhecer e delimitar o escopo da produção investigada (Guimarães; Bezerra, 2019). Um exemplo de ferramenta que complementa os estudos bibliométricos são os gerenciadores de referência, como o EndNote, que permite o processo de extração e tratamento dos dados, e o *VOSviewer*, destinado à mineração e à visualização dos dados (palavras e termos); esse último *software* gera um mapa de todas as palavras dos documentos importados, como título e resumo, com destaques às áreas ou blocos de concentração dos termos (Guimarães; Bezerra, 2019).

Ao utilizar o *VOSviewer* como ferramenta de visualização de dados para identificação das palavras que representam o tema gestão de dados, Guimarães e Bezerra (2019) descobriram que essa ferramenta pode ser utilizada na identificação das redes de colaboração entre os autores. Trata-se, portanto, de mais uma recomendação de como tratar os dados em estudos bibliométricos, tendo em vista a pluralidade de técnicas e tecnologias existentes, cujos procedimentos tendem a ampliar as análises, fornecendo novos indicadores para a produção científica (Guimarães; Bezerra, 2019). Tendo isso em foco, a próxima seção descreve esse elemento.

2.4.4 Visualização de dados

A visualização de dados ou visualização da informação é um processo de apresentar graficamente os dados, possibilitando ao usuário compreender um conjunto de dados. Por meio de recursos eletrônicos, torna-se possível aglomerar os dados, de modo visual, sem prejuízo semântico (Cairo, 2016). Assim, segundo esse autor, os recursos utilizados poderão gerar diferentes visualizações, cuja qualidade é identificada se as visualizações forem: verdadeiras (pesquisa extensiva e honesta), funcionais (representação fiel dos dados), belas (atraente e agradável), intuitivas (revelam evidências) e esclarecedoras (o público entende e aceita as evidências).

Nos estudos bibliométricos, a visualização de dados pode ser caracterizada como o mapeamento visual dos dados coletados, e é uma prática utilizada no momento da apresentação dos dados quantitativos extraídos das fontes de pesquisa e constitui uma etapa do processo de pesquisa que antecede a interpretação dos resultados. Ela potencializa as análises ao apresentar, ilustrativamente, vários aspectos ou elementos que identificam o conhecimento científico. Assim, a visualização ajuda a explorar esse conhecimento, “[...] tornando-o mais vivo, visualizando e mostrando o impacto da pesquisa, exibindo a relação estrutural e dinâmica dos domínios da pesquisa científica [...]”. Para tanto, os estudos bibliométricos requerem o desenvolvimento de várias ferramentas ou *softwares* que incorporam diferentes técnicas necessárias para garantir uma apresentação coerente, concisa, detalhada e esclarecedora (Pradhan, 2017, p. 19, tradução nossa). Desse modo, tal prática está condicionada ao uso de recursos eletrônicos formadores dos *softwares* de visualização, os quais permitem apresentar as características de alguns elementos, como: coautoria, coocorrência, citação, pares bibliográficos e cocitação (Moreira; Guimarães; Tsunoda, 2020).

Zhang e Eichmann (2019) defendem que a visualização de dados, além de uma teoria, área de pesquisa ou prática, constitui uma abordagem de pesquisa que gerencia os dados, cujo propósito é fornecer representações gráficas e ilustrativas, capazes de melhorar a compreensão do significado expresso nos dados extraídos das fontes de consulta. Em linhas gerais, de acordo com esses autores, a visualização permeia o processo de apresentação dos dados, mediante a identificação e a análise dos padrões das publicações acadêmicas. Esse estudo citou como *software* específico, que permite a visualização em estudos bibliométricos, o *CiteSpace*, capaz de modelar e computar os dados, revelando as estruturas de conhecimento (ou redes), os estudos significativos, os tópicos notáveis e as tendências de desenvolvimento na literatura sobre uma temática específica.

Dentre alguns *softwares* mencionados na literatura, Pradhan (2017) destaca a existência de 19 *softwares* que são amplamente utilizados pelos meios acadêmicos e científicos em todo o mundo. Por sua vez, Moreira, Guimarães e Tsunoda (2020) também mapearam 16 *softwares* que podem ser utilizados no processo de visualização em estudos bibliométricos. Os autores apontam que os programas oferecem aos usuários um alto grau de flexibilidade, tanto no gerenciamento quanto na análise de dados. Analisando os programas citados nos dois estudos, percebe-se que eles apresentam características em comum, em termos de estruturação dos dados, mas se diferem em termos funcionais, sobretudo na formação dos desenhos e das possibilidades de apresentação dos elementos que caracterizam a produção científica mapeada

Dentre os *softwares* analisados pelos autores, destacam-se os seguintes:

- ***BibExcel***: especificamente projetado para auxiliar um usuário na análise de dados bibliográficos, ou quaisquer dados de natureza textual, sendo que a ideia é gerar arquivos de dados que possam ser importados para o *Excel* ou qualquer programa que receba dados com registros, para posterior processamento;
- ***Biblio***: é um conjunto de *scripts* em linguagem *Python*, que transforma o material bibliográfico bruto extraído da *Web of Science* em mapas de ciência, reunindo informações relevantes sobre milhões de artigos em uma única imagem;
- ***Biblioshiny***: fornece um conjunto de ferramentas para pesquisas relacionadas à Bibliometria e à Cientometria e apresenta código aberto que abrange uma gama de análises quantitativas, que avaliam três grandes grupos analisados: 1 – autores: inclui dados acadêmicos, afiliação e países; 2 - fontes de publicações: avalia o impacto das fontes e a produtividade; e 3 – documentos: analisa referências, palavras e citações;

- **CiteSpace**: especializado na visualização de redes que possibilitam a exploração da paisagem intelectual de um campo de pesquisa ou disciplina. Permite a seleção de um ou mais tipos de nós que comporão a rede, elucidando os relacionamentos, as aproximações e os distanciamentos entre os elementos da rede;
- **Publish or Perish**: recupera e analisa as citações acadêmicas de diferentes fontes, incluindo o *Google Acadêmico* e a *Microsoft Academic Research*. Dentre os elementos analisados e apresentados graficamente, citam-se: o número de trabalhos, total de citações, média de citações por trabalho, número de citações por autor (com possibilidade de filtragem por ano), número de trabalhos por autor e média de autores por trabalho;
- **VOSviewer**: assim como o *Publish or Perish*, o *VOSviewer* se destaca por sua influência na apresentação de redes gráficas. Trata-se, portanto, de *software* especializado que possibilita a criação, visualização e exploração de mapas baseados em dados de redes. Nesse programa, as redes podem ser formadas e visualizadas, considerando os seguintes elementos de análise: coautoria, coocorrência, citação, pares bibliográficos e cocitação (Moreira; Guimarães; Tsunoda, 2020).

Mesmo com a pluralidade de *software* existente, Pradhan (2017) destaca que a escolha pode ser baseada em critérios subjetivos, pois cada um desses *softwares* tem seu próprio conjunto de recursos, funcionalidades e finalidades específicas voltadas à execução de determinadas tarefas. A maioria dessas ferramentas está disponível gratuitamente para os usuários na internet, cabendo aos pesquisadores escolherem, a partir de critérios preestabelecidos, qual ferramenta melhor atende aos propósitos do que se quer apresentar graficamente (Pradhan, 2017; Moreira; Guimarães; Tsunoda, 2020).

Dentre os *softwares* recomendados na literatura, uma funcionalidade bastante explorada é a formação das redes que, ao combinarem elementos, facilita o entendimento de informações e conhecimentos da área estudada (Moraes; Kafure, 2020). As autoras mencionam que o aperfeiçoamento dos *softwares* favorece a visualização e a exploração de aspectos gerais e específicos e, assim, aponta caminhos e abordagens possíveis para definição da abrangência e delimitação do escopo de uma pesquisa científica. Ao utilizar o *software VOSviewer* para mapear a produção científica sobre estudos de usuários, as autoras conseguiram “[...] construir, visualizar e analisar as redes de co-ocorrência de palavras (em Inglês) e as redes de co-citação de autores, de referências e de fontes [...]”, cuja ferramenta se

mostrou eficaz ao localizar os interlocutores, o que eles discutem e a produção científica deles (Moraes; Kafure, 2020, p. 1).

Uma das características principais do estudo supracitado foi a formação das redes bibliométricas, com o uso do *VOSviewer*. O potencial desse *software* está na capacidade de formar e apresentar os elementos das redes e suas relações referentes a periódicos, pesquisadores ou publicações individuais. As redes podem ser construídas e visualizadas a partir da interação entre elementos, tais como: citação, acoplamento bibliográfico, cocitação ou relações de coautoria. O Quadro 10 detalha os cinco tipos de redes desenvolvidas no *VOSviewer*, conforme procedimentos adotados por Moraes e Kafure (2020).

Quadro 10 – Tipos de redes elaboradas com o uso do VOSviewer

Tipo de análise	Unidade de análise	Descrição da relação
Coautoria	Autores, organizações ou países.	A relação dos itens é determinada com base na coautoria de documentos.
Coocorrência	Todas as palavras-chave, palavras-chave do autor ou palavras-chave plus.	A relação dos itens é determinada com base no número de documentos nos quais eles ocorrem juntos.
Citação	Documentos, fontes, autores, organizações ou países.	A relação dos itens é determinada com base no número de vezes que eles se citam nos documentos.
Acoplamento bibliográfico	Documentos, fontes, autores, organizações ou países.	A relação dos itens é determinada com base no número de referências que eles compartilham nos documentos.
Cocitação	Referências citadas, fontes citadas ou autores citados (primeiro autor).	A relação dos itens é determinada com base no número de vezes que são citados em conjunto nos documentos.

Fonte: Moraes e Kafure (2020, p. 8).

Os *softwares* e as técnicas bibliométricas utilizadas no processo de visualização podem ser combinados, com o propósito de aferir mais qualidade à visualização, conforme recomendado por Moraes e Kafure (2020), que complementaram as análises com a mineração de textos. A mineração, segundo Souza e Perry (2019), faz parte do processo de análise de dados e, em conjunto com a visualização dos dados, coloca o pesquisador diante de uma grande quantidade de documentos, sendo necessário utilizar técnicas específicas para as análises, haja vista extrair informações úteis a partir de coleções de documentos.

Uma técnica de mineração bastante utilizada é a mineração por grafos, ao atribuir características tridimensionais aos termos que compõem os textos dos documentos. Nessa técnica, “[...] cada dimensão pode indicar o número de vezes que o termo aparece no texto, ou

pode indicar o peso do termo a ser considerado, como por exemplo, a quantidade de documentos nos quais o termo aparece [...]” (Souza; Perry, 2019, p. 217).

No que tange à apresentação dos resultados da mineração, Souza e Perry (2019) ensinam que os sistemas são fundamentados em rotinas de pré-processamento, algoritmos para descoberta de padrões e elementos para apresentação dos resultados, o que garante o processo de visualização dos dados a serem interpretados. Tal como ocorre na visualização, os dados submetidos à mineração de textos são processados por *softwares*, como o *Sobek*, que gera grafos e pode ser utilizado em diferentes áreas e contextos, por exemplo, na prática de representação de documentos (Medeiros; Pinho; Correa, 2020).

Assim como Medeiros, Pinho e Correa (2020) validaram a aplicação do *Sobek* no processo de mineração de textos no campo da organização e representação da informação, entende-se que tal ferramenta se estende aos estudos bibliométricos, como pontuado por Barbosa, Severo e Reategui (2009). Desse modo, de posse dos dados coletados nas fontes, utiliza-se a ferramenta que faz o processamento “[...] das palavras registradas e as ocorrências de palavras repetidas ou sinônimas no documento, fazendo relações e criando grafos de interação entre elas, expondo os principais termos/conceitos do texto em mineração [...]” (Barbosa; Severo; Reategui, 2009, p. 6).

Destaca-se, também, que a mineração de textos, ao ser conduzida por um *software* específico, precisa obedecer a uma série de procedimentos agrupados em etapas, como mencionado por Medeiros, Pinho e Correa (2020). Os autores recomendaram o uso do *software* em três etapas: 1 - organização textual do conteúdo documental; 2- inserção do conteúdo textual organizado no *software Sobek* e geração do grafo; e 3 - análise do grafo gerado pelo *software*, cujos resultados indicaram a viabilidade do *software* e da mineração.

O uso do *Sobek* no processo de mineração de textos durante as análises bibliométricas pode ser complementado com outros *softwares*, tal como o *Voyant Tools*, especializado na formação de nuvens de palavras, o que fortalece o processo de visualização dos dados. Esse procedimento foi realizado por Führl e Alvarez (2021), ao mapear a produção científica sobre o tema humanidades digitais e ciência aberta. Com esses *softwares*, os processos de análise e visualização ficam mais evidentes, sobretudo ao estabelecer redes de relacionamentos entre os elementos. Essa constatação, obtida por Führl e Alvarez (2021), reforça as múltiplas aplicações desses métodos, ao identificar com clareza o compartilhamento, a circulação, os dados, o processo colaborativo, as tecnologias, entre outras possibilidades.

2.4.5 Redes de colaboração

A expressão “redes de colaboração” pode assumir diferentes significados, a depender do contexto em que é utilizada. Na área de Ciências Sociais Aplicadas, do ponto de vista sociológico, considerando as relações sociais que projetam a sociedade, as redes manifestam-se como “[...] a nova morfologia social de nossas sociedades e a difusão da lógica de redes modifica de forma substancial a operação e os resultados dos processos produtivos e de experiência, poder e cultura” (Castells, 1999, p. 497).

O conceito clássico de redes formulado por Castells (1999) revela que a dinâmica social é movida por essas redes, ou seja, a sociedade se desenvolve por meio do vínculo estabelecido entre pessoas e grupos, mediante princípios compartilhados na tentativa de se obter resultados comuns. No contexto das transformações tecnológicas, segundo o autor, as redes se multiplicaram, como também se entrelaçaram em um circuito contínuo que fica difícil identificar claramente o início e o fim das relações. Tais redes, na perspectiva sociológica, projetam-se para diferentes contextos de vínculos, como: afetivos, de entretenimento, de negócios, comunicacionais, científicos, profissionais, dentre outros (Castells, 1999).

No caso dos vínculos científicos, surgem as redes de colaboração acadêmica, ou seja, o aglomerado de elementos que permeiam o desenvolvimento científico, incluindo pesquisadores, instituições, canais de comunicação, dentre outros elementos, que estabelecem vínculos para compartilhamento de ideias, auxílios mútuos, dentre outros benefícios compartilhados (Silva, 2018). Trata-se de um processo de integração para compartilhamento, parceria, colaboração, tendo em vista facilitar o trabalho dos envolvidos que instituem laços de relacionamento e delimitam seus campos de atuação (Silva, 2018).

Esse *locus* interativo forma-se pela congregação de indivíduos aglutinados em torno de interesses comuns, sabedores da importância do esforço coletivo e da divisão de tarefas como fundamentos propulsores das forças produtivas de pesquisa. Com isso, aumenta-se a plausibilidade de consecução dos objetivos de um conjunto de atores e o conhecimento produzido torna-se mais completo, tendo em vista que múltiplas visões foram empreendidas e a dialogia exercitada por um conjunto de mentes distintas, porém, com propósitos comuns (Sobral *et al.*, 2020, p. 2).

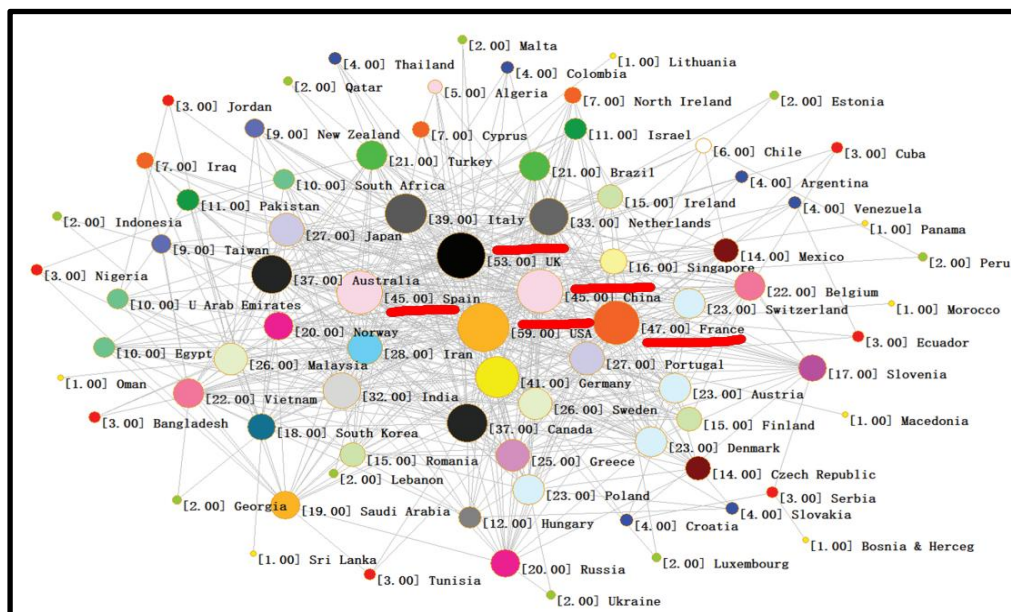
Percebe-se que o trabalho em redes é vantajoso, pois garante benefícios, como: rapidez, flexibilidade, compartilhamento de ideias, julgamentos distintos, ampliação do conhecimento, adaptação, dentre outros (Oliveira, 2018). As redes de colaboração têm como função principal o compartilhamento dos resultados das pesquisas, sejam eles parciais ou totais, com outros pesquisadores, de modo que seja possível ampliar a discussão desses resultados, “[...]”

trazendo detalhes importantes gerados por expertises em áreas interligadas, mesmo que distintas, mas que se completam [...]” (Oliveira, 2018, p. 9). Logo, presume-se que o trabalho em redes de pesquisa “[...] fornece uma aprendizagem compartilhada, novas oportunidades de investigações, o estabelecimento de novos projetos, aplicações conjuntas de fundos e transferência de tecnologia [...]” (Oliveira, 2018, p. 9). Oliveira (2021) concorda com as vantagens elencadas por Oliveira (2018) e esclarece que as redes permitem o resguardo da memória intelectual, sobretudo no caso da formação de redes genealógicas, por meio do vínculo estabelecido entre formandos e orientadores. Nessa perspectiva, esse autor profere que a análise se volta para as contribuições dos pesquisadores, tendo em vista o desenvolvimento dos campos disciplinares em que atuam, de modo a garantir a herança intelectual da área e os vários elementos produzidos com as relações estabelecidas durante o desenvolvimento científico.

As redes de colaboração são formadas a partir do entrelaçamento de elementos constituídos de uma amostra de pesquisa, tais como: tipo de documento, idioma de publicação, produção científica, publicação de periódicos, dentre outros (Sun; Yuan, 2020). Esses elementos podem ser representados em redes, considerando, por exemplo, suas autorias e a contagem das palavras-chave que representam o conteúdo dos documentos. Portanto, manifestam-se, nesses casos, as redes de coautoria, que representam autores, organizações e/ou países, e as redes de coocorrência, relacionadas à semelhança entre palavras e termos (Sun; Yuan, 2020).

Na visão de Hu, Wang e Deng (2020), no âmbito da bibliometria, as redes são representadas por disposições gráficas, geradas a partir do uso de *softwares* que coletam dados da amostra de pesquisa, em fontes, como exemplo, as bases de dados, e processam os dados, dispondo-os em uma representação ilustrativa. Com a finalidade de analisar a rede de colaboração mundial na temática Inteligência Artificial, levando em conta os autores, instituições e países (rede de coautoria), o citado estudo extraiu dados da *Web of Science*, formando um banco de dados *SQL Server*, cuja linguagem de programação foi a *Python*. Com o auxílio da *Python* e dos *softwares Matlab e Microsoft Excel*, foi identificada a rede de colaboração mundial, conforme disposto na Figura 1.

Figura 1 – Rede de colaboração entre países quanto ao tema Inteligência Artificial



Fonte: Adaptado de Hu, Wang e Deng (2020).

A Figura 1 expõe a visualização gráfica que as nações/regiões representadas pelos *nodos* com os maiores graus são os Estados Unidos, o Reino Unido, a França, a Espanha, a China e a Alemanha. Uma correlação positiva é observada entre a magnitude dos graus e a quantidade de nodos diretamente conectados a um nodo específico na rede. Isso indica uma tendência de que as nações representadas por esse nodo mantêm relações cooperativas com um número elevado de países. Notavelmente, os Estados Unidos e o Reino Unido posicionam-se no núcleo da rede, conduzindo pesquisas colaborativas na área da Inteligência Artificial com diversas nações/regiões ao redor do globo (Hu; Wang; Deng, 2020).

Na área da Ciência da Informação, em particular, nos estudos bibliométricos, é possível notar a formação das redes nos seguintes casos: coautorias na produção de artigos, citações, participação em bancas de avaliação, formações, entre outras possibilidades (Silva, 2018; Moraes; Kafure, 2020; Carvalho, 2021). Dentre essas, Moraes e Kafure (2020) destacam como mais recorrentes e projetadas para análise de *softwares* cinco tipos: coautoria, coocorrência, citação, acoplamento bibliográfico e cocitação. Esses tipos de redes comungam dos mesmos benefícios no que tange à cooperação, tanto institucional (economia de recursos, políticas públicas, dentre outros) quanto pessoal (rapidez, prestígio, dentre outros), mas apresentam diferenças quanto aos elementos constituintes das redes e o que é identificado nas relações estabelecidas. O Quadro 11 caracteriza as cinco principais modalidades de redes com menção ao local de estudo e aos autores que as analisaram.

Quadro 11 – Tipos de redes de colaboração científica nos estudos bibliométricos

Tipo	Descrição	Local do estudo	Autores
Coautoria	Esse tipo de rede representa um empreendimento que envolve metas comuns, esforço coletivo e resultados dos quais os colaboradores compartilham responsabilidades e créditos. O foco está nas autorias dos estudos que revelam um indicador de colaboração e trabalho conjunto.	Programas de Pós-Graduação em Ciência da Informação	Justino (2019)
Coocorrência	Refere-se à presença comum ou frequência de ocorrência e proximidade de palavras-chave semelhantes nos documentos.	Web of Science	Moraes e Kafure (2020)
Citação	Diz respeito ao número de citações recebidas por um documento ou o número total de citações recebidas por todos os documentos publicados por uma fonte, um autor, uma organização ou um país.	Bases de dados para informações de pesquisas: Dimensions e Altmetric Explorer	Gontijo e Araújo (2021)
Acoplamento bibliográfico	A finalidade dessa rede é medir a similaridade das fontes de documentos, utilizando-se da contagem do número de referências bibliográficas que são compartilhadas por eles. Logo, quanto mais próximos dois autores se encontrarem na rede, maior é o acoplamento entre eles, o que pode indicar núcleos referenciais de pesquisas.	Scopus	Silva <i>et al.</i> (2021)
Cocitação	Definida como a contagem das citações de dois documentos ou autores em um mesmo texto, ou seja, conta-se o par de citações como uma unidade e essa contagem determina uma potencial relação bibliográfica entre essas entidades (documentos ou autores).	Web of Science	Carvalho <i>et al.</i> (2021)

Fonte: Adaptado de Justino (2019), Moraes e Kafure (2020), Carvalho *et al.* (2021), Gontijo e Araújo (2021) e Silva *et al.* (2021).

A rede de coautoria é um tipo de rede direcionado à parceria entre dois ou mais autores na produção de estudos e sua devida publicação. A respeito desse tipo de rede, ao investigar as colaborações entre os PPG, Justino (2019) descobriu que a colaboração se faz presente, contemplando colaboração em coautorias em instância nacional e internacional. Dentre as temáticas mais estudadas com presença de colaboração internacional, destacaram-se, na sequência: Ciência da Informação, Organização do Conhecimento, Produção Científica, Estudos Bibliométricos e Bibliotecas. Também foram elencadas as seguintes temáticas na colaboração internacional: Estudos Bibliométricos, Produção Científica, Organização do Conhecimento, Informação e Ética. Esses resultados evidenciam, segundo o autor, o crescimento e consolidação da área da Ciência da Informação na realidade brasileira (Justino, 2019).

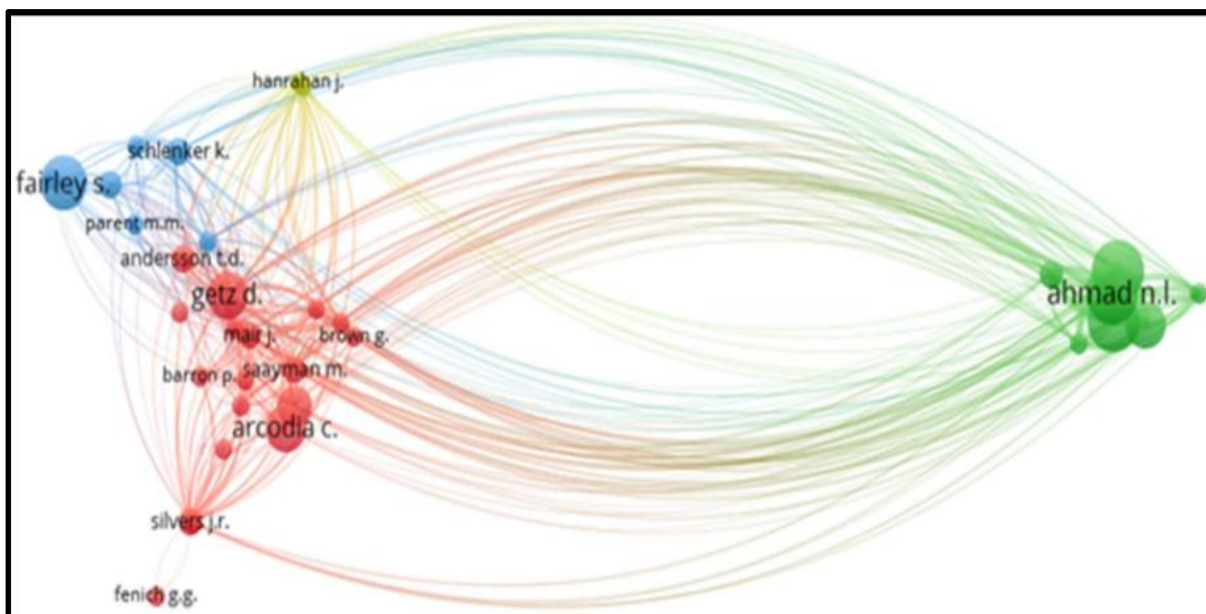
O tipo de rede voltado para o mapeamento das palavras, denominada de rede de coocorrência, serve para revelar as aproximações entre os interesses de pesquisa,

identificando os temas, seus assuntos e relações interdisciplinares (Moraes; Kafure, 2020). O estudo citado defendeu que as redes de coocorrência são geralmente usadas para fornecer uma visualização gráfica das possíveis relações entre conceitos ou entidades representadas em documentos eletrônicos. De acordo com os autores, o processo de construção de uma rede de coocorrência consiste de três etapas: primeiro, uma matriz de similaridade é calculada com base na matriz de coocorrência; segundo, a rede é construída aplicando a técnica de mapeamento à matriz de similaridade; e na terceira etapa, a rede é traduzida, girada e refletida.

A rede de citação refere-se aos estudos citados na literatura, isso é, um estudo, depois de publicado, é mencionado em outros estudos, oportunizando, assim, a interligação entre o que está sendo produzido. Sobre isso, Gontijo e Araújo (2021), mediante coleta de dados altimétricos, localizaram uma amostra de 3.121 artigos publicados entre 2014 e 2019. Quanto à rede de citações investigada, percebeu-se que o impacto acadêmico registrou cobertura de 70,81% dos artigos com dados de citações e, para a visibilidade, 61,93% obtiveram dados altimétricos referentes às menções na web social. Por isso, os autores concluíram que, “[...] ao complementar os estudos métricos tradicionais, como os bibliométricos e de citação, a altmetria [sic] e seus indicadores alternativos possibilitam uma compreensão considerada mais ampla do real impacto causado pelas produções científicas” (Gontijo; Araújo, 2021, p. 1).

O acoplamento bibliográfico refere-se às referências compartilhadas entre os documentos, podendo ligar autores, instituições, fontes de referência, dentre outros elementos, à medida que são referenciados, cujo resultado desse processo é identificar a aproximação entre os autores (Silva *et al.*, 2021). Nesse estudo, ao mapear as características da produção científica internacional sobre gestão de eventos, na Scopus, foi possível formar uma rede de acoplamento, com quatro grupos ou clusters delimitados, conforme ilustrado na Figura 2.

Figura 2 – Rede de acoplamento bibliográfico a partir de levantamento na Scopus



Fonte: Silva *et al.* (2021, p.142).

Pelas evidências do estudo de Silva *et al.* (2021), tendo como base as informações da Figura 2, constata-se que o primeiro *cluster* (vermelho) é formado por 18 autores e é composto por Austrália, Estados Unidos e Reino Unido, com predominância de instituições Australianas. O *cluster* 2 (verde), composto por autores da Malásia, apresenta-se como isolado (ponto extremo direito da rede), o que indica pouco acoplamento com os demais grupos. Ao final, os clusters 3 (azul) e 4 (amarelo) são compostos por sete e dois autores, respectivamente, e estão mais integrados ao grupo da Austrália, indicando os interesses de pesquisa e aproximações entre esses grupos.

Por fim, mas não menos importante, tem-se a rede de cocitação, ou seja, a quantidade de citações de dois documentos ou autores em um mesmo texto. Esse tipo de rede tem sido muito investigado nos estudos bibliométricos na Ciência da Informação, tal como no estudo de Carvalho *et al.* (2021). Ao extrair uma amostra da Web of Science, foi descoberto que “[...] a contagem da citação por documento citante distribui melhor os autores em um corpus de análise do que as citações totais [...]” (Carvalho *et al.*, 2021, p. 26). Ademais, conforme concluído no estudo, no que tange à cocitação do autor com ele mesmo, esse foi um indicador pouco apropriado, uma vez que não foi constatada relação direta com a influência de um autor em um conjunto de documentos, em especial nos estudos tradicionais de cocitação de autores.

2.5 Resumo do capítulo

Este capítulo discutiu os fundamentos da comunicação científica e sua relação com a estrutura e dinâmica da produção acadêmica, especialmente no contexto da Pós-Graduação no Brasil. O avanço da comunicação científica está diretamente ligado à evolução dos meios tecnológicos e metodológicos utilizados para registrar, disseminar e avaliar o conhecimento produzido. Nesse sentido, a Ciência da Informação investiga os processos de organização, recuperação e análise da informação científica, utilizando abordagens como a bibliometria para mensurar a produtividade e a colaboração entre pesquisadores e instituições.

A integração entre os fundamentos epistemológicos da Ciência da Informação e os objetos empíricos da análise permite qualificar a própria prática bibliométrica, situando-a não apenas como técnica de mensuração, mas como ferramenta analítica que opera sobre artefatos socialmente construídos. Compreender as teses e dissertações como documentos inscritos em campos discursivos, e não apenas como unidades de contagem, amplia o potencial interpretativo dos indicadores. Assim, a bibliometria pode ser inserida em uma abordagem crítica e reflexiva da comunicação científica, considerando as dinâmicas sociais, institucionais e cognitivas que moldam a produção e circulação da informação acadêmica. É nesse enquadramento ampliado que se insere a análise da produção dos Programas de Pós-Graduação em Ciência da Informação no Brasil, foco empírico deste estudo.

Além de apresentar um panorama histórico da Pós-Graduação no Brasil, o capítulo revisitou a consolidação da Ciência da Informação como área do conhecimento, destacando sua contribuição para o estudo da comunicação científica. Foram explorados os métodos bibliométricos aplicados à avaliação da produção acadêmica, evidenciando como esses instrumentos permitem mapear padrões e tendências na ciência. A literatura sobre o assunto é constituída por muitos estudos que se aprofundaram no campo da comunicação científica, com uso das técnicas e métodos bibliométricos. Embora haja essa evidência, há algumas limitações ainda não resolvidas, tal como a produtividade dos PPG da Ciência da Informação brasileira, abordagem que constitui o foco principal deste estudo.

A revisão teórica também considerou desafios contemporâneos, como a necessidade de maior transparência e acessibilidade dos dados científicos. Os estudos bibliométricos dependem diretamente da qualidade dos dados disponibilizados, uma vez que a precisão das análises está vinculada à integridade, consistência e acessibilidade das informações utilizadas. Para que esses dados sejam efetivamente exploráveis, é necessário que atendam a padrões que garantam sua encontrabilidade, acessibilidade, interoperabilidade e reutilização, conforme

preconizado pelos princípios *FAIR* (*Findable, Accessible, Interoperable, Reusable*). Embora esta tese não tenha abordado diretamente a qualidade dos dados da CAPES, essa questão foi discutida em um artigo do autor (Marques; Maculan; Souza, 2024), no qual são analisados os desafios e as oportunidades para a adoção dos princípios FAIR no contexto dos dados acadêmicos.

Complementarmente, uma integração mais densa entre os conceitos estruturantes da área e os eixos analíticos do estudo permite expandir a compreensão da produção acadêmica como parte de um campo informacional dinâmico. A partir do paradigma social de Capurro (2003), os processos informacionais são compreendidos como construções coletivas inseridas em contextos socioculturais específicos, o que confere às teses e dissertações um papel ativo na mediação entre sujeitos, saberes e instituições. Essa concepção se articula com a definição de documento como entidade social proposta por Le Coadic (1996), que reconhece no documento uma materialização intencional de práticas e sentidos. Wersig (1979), por sua vez, enfatiza que os programas de pós-graduação operam como estruturas de institucionalização do conhecimento, organizando fluxos informacionais que conectam formação, pesquisa e comunicação científica. Já o modelo sistêmico de Choo (1998) permite interpretar os dados extraídos dessas produções como artefatos informacionais que se inserem em ciclos de construção de sentido, aprendizagem e tomada de decisão. Ao articular essas abordagens, este capítulo reforça a contribuição da Ciência da Informação para a análise crítica dos mecanismos que estruturam a produção acadêmica e sua circulação nos sistemas sociais de informação.

As discussões apresentadas fornecem, assim, a base conceitual necessária para compreender a comunicação científica e sua relação com a produtividade e colaboração nos Programas de Pós-Graduação em Ciência da Informação. Essa fundamentação permite delinear os principais conceitos e identificar as técnicas bibliométricas aplicáveis à análise da produtividade acadêmica. Como desdobramento, o capítulo seguinte apresenta uma revisão integrativa da literatura, cujo objetivo é mapear os estudos já realizados sobre a pós-graduação brasileira — com ênfase na Ciência da Informação — e identificar abordagens, lacunas e contribuições relevantes que subsidiam o desenvolvimento da presente investigação.

3 REVISÃO INTEGRATIVA DA LITERATURA: ESTUDOS BIBLIOMÉTRICOS DA PRODUÇÃO CIENTÍFICA NA CIÊNCIA DA INFORMAÇÃO

Dando continuidade à fundamentação teórica apresentada no capítulo anterior, este capítulo apresenta os estudos correlatos que tratam de abordagens bibliométricas aplicadas à produção científica da área da Ciência da Informação. Por meio de uma revisão integrativa da literatura, buscou-se mapear o estado da arte sobre a análise da pós-graduação brasileira, com ênfase especial na Ciência da Informação.

A revisão enfoca, prioritariamente, estudos que têm como objeto de análise as teses e dissertações produzidas nos programas de pós-graduação, embora também considere pesquisas de áreas afins que abordam temáticas convergentes. A adoção da revisão integrativa como método permitiu a identificação, análise e síntese de pesquisas publicadas em diversas fontes, contribuindo para o reconhecimento de avanços consolidados e lacunas ainda presentes no campo. O capítulo descreve detalhadamente as etapas do processo metodológico da revisão, como a definição dos critérios de inclusão e exclusão, a seleção das bases de dados, as estratégias de busca e os procedimentos de categorização e análise dos estudos selecionados.

A discussão dos resultados permite evidenciar os enfoques predominantes, os métodos recorrentes e os desafios mais citados pela literatura, o que possibilita situar a presente pesquisa no cenário científico da área. Ao sistematizar o conhecimento existente, a revisão reforça a escassez de estudos que articulem curadoria de dados abertos e análises bibliométricas aplicadas à produção de teses e dissertações, especialmente na área da Ciência da Informação — lacuna que justifica e destaca a originalidade da proposta desenvolvida neste trabalho.

3.1 Procedimentos da Revisão integrativa da literatura (RIL)

A revisão integrativa se baseou no modelo proposto por Botelho, Cunha e Macedo (2011) para esse tipo de revisão. Os autores definem a revisão integrativa como método de pesquisa baseado no levantamento de fontes bibliográficas com o objetivo de “[...] traçar uma análise sobre o conhecimento já construído em pesquisas anteriores sobre um determinado tema” (Botelho; Cunha; Macedo, 2011, p. 127).

Para a elaboração de revisões integrativas, segundo os autores, é preciso dividir o percurso metodológico em cinco etapas, tendo em vista garantir a integração entre as

atividades realizadas durante todo o processo de revisão. As etapas definidas pelos autores e suas principais características e atividades são:

Primeira etapa: identificação do tema e seleção da questão de pesquisa (da revisão): Nesta etapa, são definidos os seguintes elementos: a) a questão a ser respondida, formulada de modo claro e específico, e alinhada ao problema de pesquisa; b) as palavras-chave a serem utilizadas durante as estratégias de busca; e c) a escolha das bases de dados a serem consultadas.

Segunda etapa: estabelecimento dos critérios de inclusão e exclusão: É nesse momento que são estabelecidos os parâmetros delimitadores que determinam os estudos componentes da revisão, justificando o que pode ser selecionado e o que pode ser eliminado. À medida que a estratégia de busca é aplicada, deve-se rever a funcionalidade dos critérios para responder à questão proposta.

Terceira etapa: identificação dos estudos pré-selecionados e selecionados: Nesta etapa, a princípio, é realizada uma análise inicial para pré-seleção da amostra, com a leitura dos seguintes elementos dos estudos recuperados: títulos, resumos e palavras-chave. Em seguida, para os casos dos estudos em que a análise desses elementos não seja suficiente para a pré-seleção, é realizada uma busca no texto completo. Finalizada a análise, é elaborada uma tabela com a seleção dos estudos que atenderam aos critérios de inclusão e exclusão, estudos esses que comporão a amostra de análise da revisão integrativa.

Quarta etapa: categorização dos estudos selecionados: Nesta etapa, propõe-se a elaboração de uma tabela contendo as principais informações que serão extraídas dos estudos selecionados. Essas informações estarão contidas em categorias de análise que irão sistematizar a síntese dos estudos. Botelho, Cunha e Macedo (2011) citam como exemplo de categoria, o título das publicações e seus autores, recomendando que outras informações que caracterizam os estudos podem também ser apresentadas.

Quinta etapa: análise e interpretação dos resultados: Nesta etapa, após a formação da amostra, os estudos são individualizados e submetidos a uma análise interpretativa e crítica dos dados coletados. Nessa atividade, o propósito é identificar as lacunas encontradas na literatura e reconhecer a necessidade de se desenvolver futuras pesquisas.

As cinco etapas mencionadas serviram como orientação do processo de revisão, estando uma integrada à outra, como um contínuo, o que garante a sistematização do processo, conforme reforçado por Botelho, Cunha e Macedo (2011). Seguindo esse modelo de revisão, as atividades de cada etapa foram aplicadas, cujos resultados são apresentados na sequência.

3.2 Resultados da RIL

Os resultados respondem a cada uma das cinco etapas dos procedimentos da RIL.

3.2.1 Resultado da Primeira etapa da RIL

A partir da leitura de textos introdutórios foi possível delimitar a temática e, conseqüentemente, identificar lacunas ainda não resolvidas, o que culminou na formação da questão a ser respondida no final da revisão. Do mesmo modo, foram selecionadas as palavras-chave a serem utilizadas na estratégia de busca. Elas foram selecionadas a partir de uma primeira exploração da literatura, sendo extraídas as palavras-chave atribuídas pelos autores dos trabalhos recuperados. Por fim, as palavras-chave foram ligadas com operadores booleanos, na intenção de informar ao sistema de busca a combinação dos termos, visando a recuperar estudos com mais precisão. Essa combinação resultou na *string*⁷ (ou expressão) de busca geral, a ser utilizada nos buscadores de cada base de dados. Ressalta-se que, em cada base, a *string* de busca foi reformulada para adequação ao sistema de busca, atividade essa realizada no início da terceira etapa. Os resultados iniciais da primeira etapa são apresentados no Quadro 12.

Quadro 12 – Resultados iniciais da primeira etapa da revisão integrativa

Elemento formulado	Descrição
Questão:	Como foram desenvolvidos os estudos que tratam sobre mapeamentos bibliométricos da produção científica brasileira a partir de conteúdos de teses e dissertações?
Lista de palavras-chave (ordem alfabética):	Bibliometria; Estudo bibliométrico; Mapeamento; Produção acadêmica; Produção científica; Programa de Pós-Graduação; Rede de colaboração científica; Tese e dissertação. [e suas variações no plural]
<i>String</i> de busca geral:	“Bibliometria” OR “Estudo bibliométrico” OR “Mapeamento” AND “Produção acadêmica” OR “Produção científica” AND “Programa de Pós-Graduação” AND “Rede de colaboração científica” AND “Tese e dissertação”. [e suas variações no plural]

Fonte: elaborado pelo autor (2022).

⁷ Neste estudo adota-se o termo *string* de busca por sua precisão técnica, considerando que se refere à formulação estruturada de operadores e termos utilizada nas bases de dados.

Além da questão da revisão, das palavras-chave e da *string* de busca geral, a primeira etapa teve como resultado a escolha das fontes de busca para recuperação dos estudos. Foram selecionadas bases de dados de escopo específico à área da Biblioteconomia e Ciência da Informação e de escopo multidisciplinar. Justificam-se as escolhas de fontes de escopo multidisciplinar pela expectativa de serem recuperados estudos publicados em áreas que dialogam com a Ciência da Informação, o que pode trazer um número maior de pesquisas.

O Quadro 13 apresenta os resultados referentes às sete bases de dados selecionadas, com menção ao nome da base, uma breve descrição, o seu escopo e o respectivo endereço eletrônico para consulta.

Quadro 13 – Informações específicas das bases de dados consultadas na revisão

Base de dados	Descrição	Escopo	Endereço para consulta
1. Base de Dados Referenciais de Artigos de Periódicos em Ciência da Informação (BRAPCI)	Mantida pela Universidade Federal do Paraná, essa base identifica os títulos de periódicos da área de Ciência da Informação e indexa os seus artigos. Disponibiliza referências e resumos de 19.255 textos publicados em 57 periódicos nacionais impressos e eletrônicos da área de Ciência da Informação (BRAPCI, 2022).	Biblioteconomia e Ciência da Informação	https://brapci.inf.br/index.php/res/
2. E-Prints in Library and Information Science (E-LIS)	Base de acesso aberto para trabalhos acadêmicos em Biblioteconomia e Ciência da Informação. Mais de 12.000 documentos foram arquivados (E-LIS, 2022).	Biblioteconomia e Ciência da Informação	http://eprints.rclis.org/
3. Library, Information Science & Technology Abstracts (LISTA)	É uma base de dados de pesquisa gratuita para estudos de Biblioteconomia e Ciência da informação. A LISTA fornece indexação e abstração para centenas de periódicos, livros e relatórios de pesquisa (LISTA, 2022).	Biblioteconomia e Ciência da Informação	https://web.s.ebscohost.com/ehost/search/basic?vid=0&sid=3871fd2e-b5f6-443d-8e6f-f8ce5efbbc00%40redis
4. Information Science & Technology Abstracts (ISTA)	É o mais antigo banco de dados de Ciência da Informação produzido continuamente. Ele fornece indexação e resumos de artigos, livros, relatórios de pesquisa e anais de conferências, além de cobertura contínua dos periódicos mais importantes da área (ISTA, 2022).	Biblioteconomia e Ciência da Informação	https://search.ebscohost.com/login.aspx?authtype=ip,uid&profile=ehost&defaultdb=izh

Base de dados	Descrição	Escopo	Endereço para consulta
5. Library and Information Science Abstracts (LISA)	Indexa mais de 500 periódicos científicos, além de livros e relatórios de pesquisas. Essa base de dados também inclui o texto completo de mais de 240 periódicos científicos (LISA, 2022).	Biblioteconomia e Ciência da Informação	https://www.proquest.com/
6. Web of Science (WOS)	Base que indexa somente os periódicos mais citados em suas respectivas áreas. Possui hoje mais de 9.000 periódicos indexados (WOS, 2022).	Multidisciplinar	https://www-webofscience.ez27.periodicos.capes.gov.br/wos/woscc/basic-search
7. SCOPUS	Permite uma visão ampla de tudo que está sendo publicado cientificamente sobre um tema. A base reúne informações de 23.000 títulos de revistas de mais de 5.000 editoras de todo o mundo, além de patentes, trabalhos de eventos e e-books (SCOPUS, 2022).	Multidisciplinar	https://www-scopus.ez27.periodicos.capes.gov.br/search/form.uri?display=basic#basic

Fonte: elaborado pelo autor (2022).

Assim, conforme mostra o Quadro 13, foram selecionadas cinco fontes de escopo específico e duas fontes de escopo multidisciplinar. Em síntese, definidas a questão de revisão, a lista de palavras-chave, a *string* de busca e as sete fontes de dados, finaliza-se a primeira etapa e passa-se para a segunda etapa, da definição dos critérios de inclusão e exclusão.

3.2.2 Resultado da Segunda etapa da RIL

Nesta etapa houve a definição dos critérios de inclusão e exclusão de estudos na amostra analisada, que levou em conta as características extrínsecas dos documentos indexados (dados de forma) e as características intrínsecas (dados de conteúdo). Para a seleção dos dados de forma, tais como ano de publicação, tipo de documento e língua, foram considerados os metadados registrados nos documentos e o processo de seleção ocorreu por meio do uso de filtros disponibilizados nas plataformas de busca das próprias bases. Já os dados de conteúdo, tais como acesso ao texto completo, estudo bibliométrico e foco na análise de teses e dissertações, foram identificados por meio da leitura do título e do resumo das publicações.

Seguindo as recomendações de Botelho, Cunha e Macedo (2011), os critérios foram aplicados em todos os documentos recuperados nas fontes selecionadas, desde o momento da

aplicação da estratégia de busca. O Quadro 14 apresenta os critérios de inclusão e exclusão, com uma breve descrição sobre cada um.

Quadro 14 – Estabelecimento dos critérios de inclusão e exclusão para seleção dos estudos

Critério		Descrição	Inclusão	Exclusão
Dados de forma	Tipo de documento	Artigo de periódico indexado	x	
		Artigo publicado em anais de eventos	x	
		Teses e dissertações	x	
		Livros	x	
		Resenhas, resenhas e editoriais		x
		Outros tipos de manuscritos		x
	Língua	Português	x	
		Inglês	x	
		Outros idiomas		x
	Data da publicação (ano)	Atemporal	x	
Dados de conteúdo	Acesso ao texto	Texto completo	x	
		Partes do texto		x
	Estudo bibliométrico	Aplicou técnicas bibliométricas	x	
		Não aplicou técnicas bibliométricas		x
	Foco temático	Produção dos PPG (análise de teses e dissertações)	x	
		Produção dos PPG (não analisa teses e dissertações ou as analisa sob um viés secundário)		x

Fonte: elaborado pelo autor (2022).

O Quadro 14 indica a existência de seis critérios estabelecidos para garantir uma seleção mais apurada, que traga estudos relevantes, em uma quantidade abrangente, que seja suficiente para responder a questão proposta na revisão integrativa. Importante reforçar que as palavras-chave, embora não foram descritas como critérios, elas também funcionam como filtros, devendo aparecer no título ou no resumo das publicações, no momento das buscas nas bases de dados.

3.2.3 Resultado da Terceira etapa da RIL

Para a identificação dos estudos pré-selecionados e selecionados foi necessário, antes de iniciar as buscas propriamente ditas, adequar a *string* de busca geral, considerando as peculiaridades de cada base de dados. Essa atividade foi realizada mediante o uso da *string* de busca geral em um pré-teste, que antecedeu a pré-seleção dos materiais. No pré-teste, a *string* de busca geral foi sendo editada, sobretudo quando o sistema de busca retornava resultado vazio, até que um número significativo de documentos — isto é, suficiente para avaliar a relevância, a cobertura temática e a adequação dos termos empregados — pudesse ser recuperado. Mediante a aplicação do pré-teste, uma *string* de busca específica foi definida para cada base de dados, conforme descrito no Quadro 15.

Quadro 15 – Adequação da string de busca nas bases de dados

String de busca geral:	“Bibliometria” OR “Estudo bibliométrico” OR “Mapeamento” AND “Produção acadêmica” OR “Produção científica” AND “Programas de Pós-Graduação” AND “Redes de colaboração científica” AND “Teses e dissertações”.
BRAPCI	<i>String 1</i> - Bibliometria AND Teses e dissertações.
E-LIS	<i>String 2</i> - “Estudo bibliométrico” AND “Teses e dissertações”.
LISTA	<i>String 3</i> - “Mapeamento” OR “Estudo bibliométrico” AND “Redes de colaboração científica” AND “Produção científica” OR “Produção acadêmica” AND “Programas de Pós-Graduação” AND “Teses e dissertações”.
ISTA	<i>String 4</i> - “Mapeamento” OR “Estudo bibliométrico” AND “Redes de colaboração científica” AND “Produção científica” OR “Produção acadêmica” AND “Programas de Pós-Graduação” AND “Teses e dissertações”.
LISA	<i>String 5</i> - “Estudo bibliométrico” AND “Teses e dissertações”.
WOS	<i>String 6</i> - “Mapeamento” AND “Teses e dissertações”.
SCOPUS	<i>String 7</i> - “Estudo bibliométrico” OR “Mapeamento” AND “Teses e dissertações”.

Fonte: elaborado pelo autor (2022).

Com a adequação das *strings* de busca, iniciaram-se as buscas, cujo resultado inicial foi a pré-seleção do material, mediante a análise comparativa das palavras-chave constantes na *string* de busca geral e suas correspondências nos títulos e nos resumos das publicações. Isso quer dizer que essa primeira análise não contemplou a aplicação dos critérios de inclusão. O resultado dessa atividade, com os estudos recuperados em cada base, pode ser observado no Quadro 16.

Quadro 16 – Resultados de estudos pré-selecionados – Análise inicial da terceira etapa

Base de dados	Estudos recuperados	Estudos excluídos	Estudos pré-selecionados
BRAPCI	6	0	6
E-LIS	5	1	4
ISTA	93	72	21
LISTA	93	72	21
LISA	91	50	41
WOS	26	15	11
SCOPUS	7	0	7
TOTAL	321	210	111

Fonte: elaborado pelo autor (2022).

Como apresentado no Quadro 16, o resultado final da análise aos títulos, aos resumos e às correspondências com as palavras-chave da *string* de busca geral proporcionou uma amostra provisória composta por 111 registros. Na atividade seguinte, esses registros foram submetidos aos critérios de inclusão e exclusão, considerando, apenas, os dados de forma. Nesse processo, os filtros de seleção constantes nas plataformas de algumas bases foram acionados e, para as bases sem filtros, a análise foi realizada em cada registro (documento), de modo individual. Quanto aos seis registros pré-selecionados na BRAPCI, após análise individual, percebeu-se que todos atendiam aos critérios de inclusão.

Na E-LIS, também não houve exclusão de estudos. Na ISTA, dos 21 estudos, cinco foram excluídos por estarem escritos em outros idiomas que não o inglês e o português, ocorrência essa também constatada na base LISTA (com cinco excluídos). Já na LISA, os 41 registros eram artigos de periódicos escritos em português e inglês, porém, um foi excluído por estar no formato de editorial. Por sua vez, os registros da WOS e da SCOPUS atenderam aos critérios de inclusão quanto aos dados de forma, não havendo, portanto, exclusão de registros. O Quadro 17 resume os resultados da aplicação inicial dos critérios de inclusão e exclusão.

Quadro 17 – Aplicação dos critérios de inclusão e exclusão – dados de forma

Base de dados	Estudos recuperados	Estudos excluídos	Estudos pré-selecionados
BRAPCI	6	0	6
E-LIS	4	0	4

Base de dados	Estudos recuperados	Estudos excluídos	Estudos pré-selecionados
ISTA	21	5	16
LISTA	21	5	16
LISA	41	1	40
WOS	11	0	11
SCOPUS	7	0	7
TOTAL	111	11	100

Fonte: elaborado pelo autor (2022).

Com uma amostra preliminar de 100 registros, realizou-se a última atividade da terceira etapa, com a finalização da aplicação dos critérios de inclusão e exclusão quanto aos dados de conteúdo. Nesse processo, os registros foram individualizados, com leitura dinâmica (exploratória) realizada no título, resumo e em algumas partes do documento, como na introdução, na metodologia e na conclusão.

Na BRAPCI, dos seis registros, dois foram excluídos, visto que as teses e dissertações foram consideradas como tema secundário, pois o foco teve-se a periódicos e outros trabalhos acadêmicos. Resultados parecidos foram encontrados nos quatro registros da base E-LIS, visto que o foco foi a identificação de redes colaborativas nas publicações em periódicos, não focando, pois, nas teses e dissertações. Analisando os 16 registros da ISTA, em comparativo com a LISTA, identificou-se que se tratavam dos mesmos registros. Isso pode ser justificado pelo fato de ambas as bases estarem sob a responsabilidade administrativa da Editora EBSCO. Portanto, os registros duplicados foram eliminados e a leitura do conteúdo dos 16 estudos revelou que eles abordavam temas desconexos ao tema desta revisão, contemplando estudos desenvolvidos na área de Meio-Ambiente, sem envolver análise em teses e dissertações. Na LISA, dos 40 estudos, 15 realizaram análise bibliométrica em periódicos ou artigos de evento e 12 não estavam disponíveis para acesso gratuito; logo, foram incluídos 13 estudos. A análise aos 11 registros da WOS mostrou que um registro estava indisponível e os demais se adentravam ao foco desta revisão, sendo, portanto, incluídos. Por fim, os sete registros da SCOPUS indicaram que um estudo não focou nas teses e dissertações e outro estava indisponível para acesso gratuito, possibilitando a inclusão de

cinco registros na amostra. O Quadro 18 resume o total de estudos selecionados com a finalização da análise aos critérios de inclusão e exclusão.

Quadro 18 – Aplicação dos critérios de inclusão/exclusão – dados de conteúdo

Base de dados	Estudos pré-selecionados	Estudos excluídos	Estudos incluídos
BRAPCI	6	2	4
E-LIS	4	4	0
ISTA	16	16	0
LISTA	16	16	0
LISA	40	27	13
WOS	11	1	10
SCOPUS	7	2	5
TOTAL	100	68	32

Fonte: elaborado pelo autor (2022).

Ao finalizar a aplicação dos critérios de inclusão e exclusão, percebeu-se que 32 estudos foram selecionados. Com o intuito de validar a legibilidade desses registros, os títulos foram agrupados em quatro planilhas, uma para cada base de dados. Uma análise sistematizada e conjunta às quatro planilhas revelou a existência de três estudos repetidos: um estudo da LISA também estava contido na planilha da BRAPCI; um estudo da WOS já estava na planilha da LISA e um estudo da SCOPUS havia sido descrito na planilha da WOS. Com isso, procedeu-se à exclusão desses registros em duplicidade, gerando uma amostra definitiva de 29 artigos elegíveis para análise. A fim de ilustrar a quantidade de registros selecionados em cada atividade da terceira etapa da revisão, apresenta-se o Quadro 19.

Quadro 19 – Formação da amostra de análise ao longo das atividades da terceira etapa

Atividades	Síntese do que foi realizado	Quantidade de estudos incluída
Busca inicial	Consultas individuais em cada base de dados com as devidas <i>strings</i> .	321 registros recuperados.
Correspondências das palavras-chave	Análise da existência das palavras-chave nos títulos e nos resumos.	111 registros pré-selecionados.
Análise dos dados de forma	Leitura dinâmica dos metadados dos registros, com uso dos filtros de seleção das bases.	100 registros pré-selecionados.
Análise do conteúdo textual	Leitura dinâmica dos resumos e demais partes dos registros.	32 registros selecionados.

Atividades	Síntese do que foi realizado	Quantidade de estudos incluída
Elegibilidade dos estudos	Agrupamento dos estudos em planilhas, por base de dados, e análise comparativa, para identificar duplicidades e inconsistências.	29 registros selecionados e validados.
Amostra final definitiva		29 registros válidos

Fonte: elaborado pelo autor (2022).

Cumprindo as recomendações de Botelho, Cunha e Macedo (2011), os 29 registros componentes da amostra de pesquisa foram descritos em uma única planilha. Além da organização dos registros, tendo em vista facilitar as atividades de análise requeridas na última etapa, a sistematização dos registros na planilha possibilita revelar as características principais dos estudos, evidenciando o perfil bibliométrico da amostra de análise, conforme resultados oriundos com a aplicação da quarta etapa, apresentados a seguir.

3.2.4 Resultado da Quarta etapa da RIL

A sistematização dos dados referentes a cada estudo selecionado foi realizada por meio da descrição de categorias, conforme sugerido por Botelho, Cunha e Macedo (2011), sendo que esses dados foram considerados como categorias de análise e foram inseridos como componentes da planilha, apresentados na forma de colunas. Esses dados ou categorias fazem referência às características de cada registro, sejam elas de natureza extrínseca (forma), como título, autores, ano, dentre outras, como de natureza intrínseca (conteúdo), por exemplo: método, objetivo e outras.

Seguindo essa recomendação, foram extraídos do conjunto de textos da amostra de análise as seguintes categorias analíticas: título, autores, objetivo e conclusão, método aplicado (técnicas bibliométricas e *softwares*), visualização de dados, redes de colaboração e data de publicação (ano). O quadro síntese dos estudos analisados encontra-se no Apêndice A (Quadro 1Quadro 106), apresentando os principais atributos dos trabalhos revisados. Ele sistematiza o conjunto dessas informações, revelando os componentes da amostra de pesquisa desta revisão, apresentados em ordem cronológica. No total, foram analisados 29 estudos, a maioria de natureza bibliométrica quantitativa, com forte concentração em abordagens descritivas.

Após essa sistematização, iniciaram-se as atividades da quinta e última etapa, que se adentrou ao processo de análise profunda dos conteúdos dos textos selecionados. Para isso, os textos foram comparados, com descrição de seus principais resultados e conclusões e o que

eles indicam para a proposição de lacunas ainda não resolvidas, ou seja, a constituição de estudos futuros. Essa sistematização constituiu uma base sólida para a análise crítica realizada na etapa seguinte, orientando a identificação das lacunas teóricas e metodológicas que fundamentaram a proposição de novos estudos no campo da Ciência da Informação.

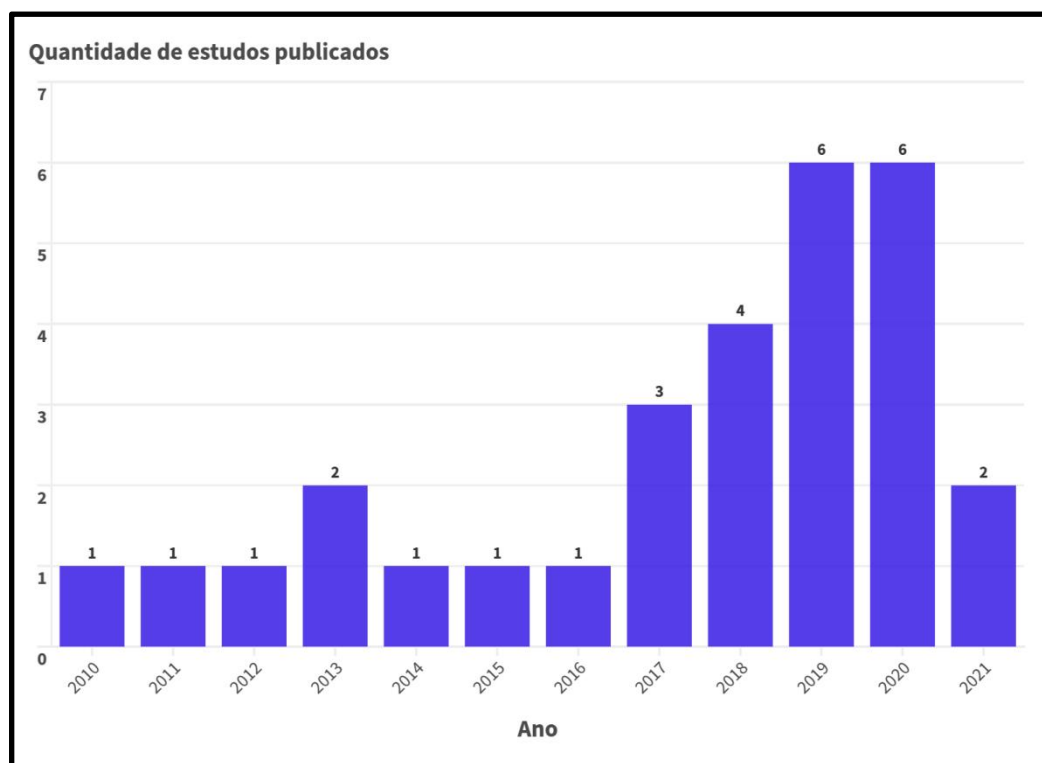
Com base nessa sistematização, a quinta etapa da revisão integrativa foi conduzida com foco na análise crítica e interpretativa dos conteúdos dos estudos selecionados, visando identificar lacunas persistentes e delinear oportunidades para pesquisas futuras no campo da Ciência da Informação.

3.2.5 Resultados da Quinta etapa da RIL

Antes da análise profunda aos conteúdos de cada estudo componente da amostra, são apresentados os resultados referentes à quantidade de autores por estudo e ao ano de publicação, esse último possibilitou constituir uma linha do tempo quanto ao percurso de desenvolvimento dos estudos da amostra. Em seguida, apresentam-se os resultados para as categorias: visualização de dados e redes de colaboração. Ao final, têm-se a exposição e a interpretação quanto às temáticas, aos objetivos/conclusão e aos métodos adotados em cada estudo. Quanto à quantidade de autores, a maioria dos estudos foi desenvolvida de modo colaborativo, com predominância de estudos com mais de quatro autores, sendo que apenas um estudo apresentou autoria individual. Esse resultado confirma a preferência pela autoria coletiva nos estudos bibliométricos.

Segundo Stueber e Teixeira (2019), a autoria coletiva ou múltipla é um ponto forte desses estudos e estimula a formação das redes de colaboração, cujos pesquisadores procuram identificar outros colegas com interesses de pesquisa similares, tanto em grupos de estudo, em linhas de pesquisa e/ou em publicações, o que fortalece a formação dessas redes. Pela verificação aos anos de publicação dos 29 estudos, foi possível estabelecer o seguinte intervalo: de 2010 a 2021. Por meio desse intervalo, constata-se que o interesse dos pesquisadores pela temática é recente, específico da segunda década do século XXI. A quantidade de estudos publicados por ano pode ser visualizada no Gráfico 3.

Gráfico 3 – Distribuição anual dos estudos selecionados para a revisão



Fonte: Dados da pesquisa (2022).

Nota-se que houve publicações em todos os anos dentro do intervalo de 2010 a 2021 e que a quantidade de estudos foi se intensificando, ao longo do tempo, exceto no último ano. Os anos de 2018, 2019 e 2020 destacam-se por possuírem maior número de estudos publicados. Presume-se que o ano de 2021 tenha apresentado uma queda (dois estudos), em virtude da paralisação nas atividades de pesquisa ocorrida no ano anterior, em decorrência da crise provocada pela pandemia da Covid-19 (Assunção-Luiz, 2021). O ano de 2022, que não apresentou nenhum estudo, também pode estar entre os efeitos da crise, além do fato de que, no momento da coleta de dados, o presente ano não havia terminado.

A análise dos anos de publicação possibilita aferir o resultado de que os estudos bibliométricos sobre a produtividade dos Programas de Pós-Graduação (PPG) brasileiros, por meio da verificação em teses e dissertações, têm crescido, ou seja, a temática tem despertado interesse na comunidade científica. Esse resultado reforça as evidências da literatura, tal como em Trindade, Carvalho e Paletta (2021) que, além de identificarem o crescimento da produção científica, associaram-no aos investimentos em projetos científicos, grupos de estudos e eventos. Assim, pode-se atestar que esses estudos, nas últimas décadas, têm acompanhado a expansão da ciência e da tecnologia, tornando-se, “[...] cada vez mais evidente a necessidade

de avaliar tais avanços e de determinar os desenvolvimentos alcançados pelas diversas disciplinas do conhecimento [...]” (Vanti, 2002, p. 152).

Analisando, em profundidade, o conteúdo integral dos estudos, foi possível identificar as estratégias adotadas para garantir a visualização dos dados bibliométricos mapeados. Todos os estudos adotaram algum recurso que tornou os dados mais apresentáveis, possibilitando melhor exposição visual. Houve predominância no uso de quadros, tabelas e gráficos e menor incidência quanto ao uso de figuras, mapas conceituais e nuvens de palavras. Observa-se que ainda há carências quanto ao uso de mapas conceituais, embora se apresentem como “[...] uma forma de organizar, representar, visualizar e compartilhar o conhecimento” (Rodrigues; Servantes, 2013, p. 770). Do mesmo modo, as nuvens de palavras são recursos de visualização de dados bibliométricos, pois facilitam a disposição gráfica e visual das palavras com maior frequência em um documento ou domínio do conhecimento (Ferreira; Silva, 2019).

Outra categoria analítica diz respeito à identificação ou formação das redes de colaboração. Dentre os 29 estudos, apenas dois mencionaram essas redes. No estudo de Stueber e Teixeira (2019), foram mapeados seis tipos de redes, a saber: coautoria, grupos de pesquisa, instituições de pesquisa, diretórios de grupos de estudo, rede intrainstitucional e redes de domínios. Em Costa e Rodrigues (2019), a rede de autores que investigam uma determinada temática foi elaborada, considerando as instituições e sua distribuição por regiões e estados brasileiros, o que possibilitou o mapeamento geográfico do tema. Esse resultado está de acordo com o estudo de Gama, Ferreira Neto e Santos (2021), pois os autores destacaram que as redes ilustram as relações entre pesquisadores, temas e instituições de diferentes locais, revelando o entrosamento desses agentes e as possibilidades de se constituírem parcerias na comunicação científica.

No que tange às temáticas dos estudos, ou seja, qual objeto foi investigado e em que área ou campo do conhecimento ele se insere, foi possível observar uma multiplicidade de ocorrências, o que confirma o uso da Bibliometria em diferentes áreas, além da natureza interdisciplinar que os métodos bibliométricos vêm se apoiando e se expandindo. Mediante a identificação das temáticas, agruparam-se os estudos por grande área, tendo em vista a relação direta do tema com disciplinas específicas dessas áreas, como também por campos do conhecimento a que os temas estão vinculados. Destaca-se que, neste agrupamento, não foi levado em consideração o PPG que realizou a pesquisa, mas a grande área e o campo do conhecimento a que os temas estão vinculados, diretamente, considerando a divisão do conhecimento científico sugerida pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). O Quadro 20 apresenta esse agrupamento.

Quadro 20 – Distribuição das temáticas/objetos dos estudos por áreas e campos do conhecimento, com autores e ano de publicação

Grande área	Subárea ou campo do conhecimento	Temática/Objeto	Autores e anos
Ciências da Saúde	Psicologia	Avaliação Psicológica	Joly <i>et al.</i> (2010)
		Gênero e Sexualidade	Rios e Souza (2017)
	Enfermagem	Educação em Enfermagem	Backes <i>et al.</i> (2013)
		Fenomenologia Social de Alfred Schütz	Vieira <i>et al.</i> (2013)
		Prática da Enfermagem em Portugal	Baggio <i>et al.</i> (2014)
		Teoria das Representações Sociais	Oliveira, Maia e Miranda (2016)
		Classificação Internacional para a Prática de Enfermagem	Beserra <i>et al.</i> (2018)
		Modelo Calgary de Avaliação Familiar	Costa <i>et al.</i> (2018)
		Serviço de Atendimento Móvel de Urgência	Agra <i>et al.</i> (2018)
	Medicina	Cuidados Paliativos em Oncologia	Dias <i>et al.</i> (2020)
		Cuidados Paliativos em Doenças Crônicas	Paiva <i>et al.</i> (2020)
		Cuidados Paliativos na Atenção Primária à Saúde	Lima <i>et al.</i> (2021)
	Educação Física	Gestão do Esporte	Moraes, Amaral e Bastos (2021)
Ciências Sociais Aplicadas	Ciência da Informação	Bibliometria	Araújo e Alvarenga (2011)
		Produção Bibliométrica de Grupo de Pesquisa	Soares, Maculan e Lima (2015)
		Comunicação Científica	Stueber e Teixeira (2019)
		Taxonomia Corporativa	Lopes, Aganette e Maculan (2020)
	Educação	Educação Indígena	Maroldi <i>et al.</i> (2017)
		Objetos de Aprendizagem	Uehara <i>et al.</i> (2017)

Grande área	Subárea ou campo do conhecimento	Temática/Objeto	Autores e anos
		Ludicidade	Santos e Pereira (2019)
		Pedagogia da Alternância	Costa e Rodrigues (2019)
		Pesquisa Educacional	Felicetti <i>et al.</i> (2019)
		Educação Ambiental	Motin <i>et al.</i> (2019)
		Currículo e Educação Especial no/do Campo	Silva e Peternella (2020)
		Letramento	Sousa (2020)
Ciências Biológicas	Genética	Transgênicos	Hayashi, Rigolin e Hayashi (2012)
Ciências Humanas	História	Protestantismo, História e Cultura Escrita	Júnior e Galvão (2018)
Ciências Agrárias	Agricultura Familiar	Pluriatividade	Silva, Fiúza e Pinto (2019)
Linguística, Letras e Artes	Artes	Moda	Trindade, Carvalho e Paletta (2021)

Fonte: elaborado pelo autor (2022).

Os dados expostos no Quadro 20 revelam a aplicação da Bibliometria em temáticas pertencentes a variadas áreas do conhecimento, com destaque a área das Ciências da Saúde, em específico o campo da Enfermagem e da Medicina, e a área das Ciências Sociais Aplicadas, predominando os campos da Educação e da Ciência da Informação. Embora tenha surgido no campo da Ciência da Informação, direcionada a quantificar os itens que compõem os acervos das bibliotecas (Fonseca, 1986), a Bibliometria vem se expandindo, visto que possibilita quantificar e avaliar a produção temática de diferentes áreas (Lima; Ferreira; Do Prado, 2022), como pode ser atestado no Quadro 10. Ao serem utilizadas em diferentes áreas, as técnicas revestem-se de um aparato interdisciplinar, recebendo contribuições teóricas e metodológicas oriundas de diferentes áreas (Rodrigues; Faiad, 2019).

Os temas da área das Ciências da Saúde também foram identificados com maior predominância no estudo de Rodrigues e Faiad (2019), com uma quantidade mais abrangente de estudos empíricos realizados, sobretudo, por profissionais da Enfermagem. Do mesmo modo, os pesquisadores da Educação têm mostrado bastante interesse na realização dos estudos bibliométricos, com a expectativa de avaliar a qualidade do ensino, “[...] sendo interessante identificar tendências e a evolução das publicações [...]” (Ribeiro *et al.*, 2019, p.

311). Assim como a Educação, o campo da Ciência da Informação tem expandido as estratégias para quantificar a produção científica, cujo aumento pode estar associado ao fato de que esse campo busca atender as necessidades sociais (Gabriel Junior; Bochi; Moura, 2021). Analisando, ainda, o Quadro 10, percebe-se que, das oito grandes áreas elaboradas pelo CNPq para a organização do conhecimento científico, apenas duas áreas não possuem temáticas analisadas, bibliométricamente: Engenharias e a área das Ciências Exatas e da Terra.

Esse resultado se aproxima do que descobriu Amaral *et al.* (2019), quando esses autores descreveram que os estudos bibliométricos realizados pelas Engenharias ainda são escassos, embora precisam ser aprofundados para garantir a qualidade da produção científica dessa área. Mesmo que as temáticas da área de Exatas, por exemplo, tenham sido abordadas pela Bibliometria, de modo tímido, há evidências de crescimento, com diferentes abordagens, tal como a análise do gênero dos pesquisadores, como investigado por Tuesta *et al.* (2019).

Finalizada a análise das temáticas da amostra de pesquisa, analisaram-se os objetivos/conclusão dos estudos. Comparando os objetivos, nota-se que eles se direcionam à aplicação das técnicas de medição sobre um ou mais aspectos que caracterizam, quantificam ou avaliam uma área do conhecimento. Mesmo com esse sentido, os textos dos objetivos apresentam sintáticas diferentes, a maioria (15 estudos) contendo a expressão “produção científica” e suas variantes, como: “produção acadêmica”, “produção brasileira”, dentre outras. Também são mencionadas expressões do tipo: “estado da arte”, “temáticas e métodos”, “constituição do campo científico”, “indicadores bibliométricos”, dentre outras ocorrências.

Quanto às conclusões, elas refletem os próprios objetivos. Assim, é descrito o resultado advindo a partir do objetivo definido, com menção às seguintes expressões referentes ao aspecto analisado: “temáticas identificadas”, “áreas vinculadas”, “autores e membros de grupos”, “quantidade de estudos”, “tendências e interesse de pesquisa” e “panorama temático”. Em algumas conclusões, é mencionada, além do resultado principal alcançado, a contribuição do estudo. Nesse sentido, destacam-se como principais expressões referentes ao que foi alcançado: “avanço do conhecimento”, “produção do conhecimento” e “aplicações”.

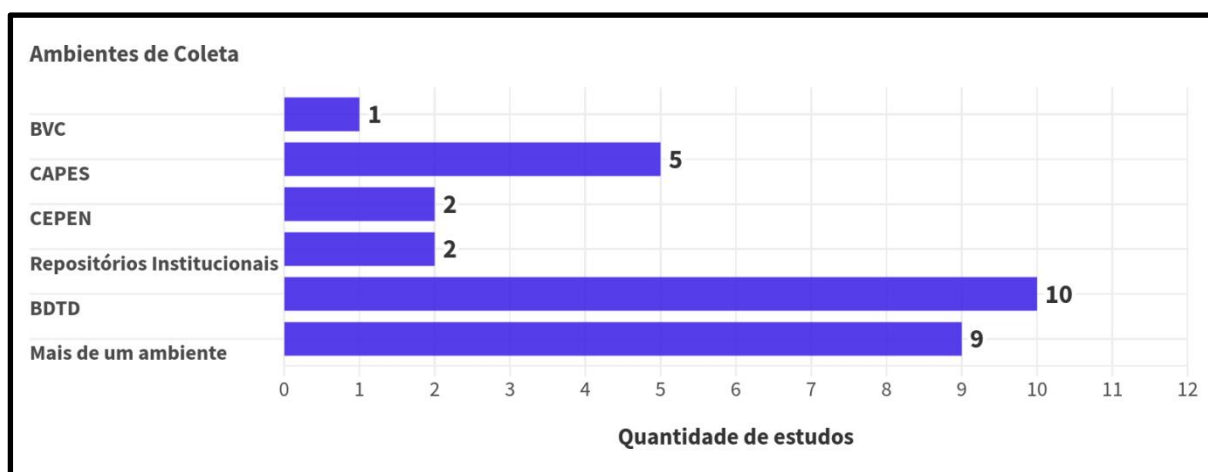
Pelo texto dos objetivos e das conclusões, confirma-se a essência dos estudos bibliométricos que é “[...] medir índices de produção e disseminação do conhecimento, bem como acompanhar o desenvolvimento de diversas áreas científicas [...]” (Costa; Lopes; Fernández-Llimós, 2012, p. 1). Ademais, os estudos que aplicam técnicas bibliométricas levam em consideração a publicação de documentos científicos, com a finalidade de mapearem “[...] o corpo de conhecimento produzido por pesquisadores especializados em

determinada área do saber [...]” (Sousa; Fontenele, 2019, p. 214). O resultado desse processo será, em sua essência, “[...] avaliar a produtividade e a qualidade da pesquisa dos cientistas [...]” (Machado Junior *et al.*, 2016, p. 116), tendo em vista a proposição de melhorias para as áreas, por conseguinte, oportunizar o desenvolvimento científico e tecnológico da ciência e da nação (Medeiros; Vitoriano, 2015).

Por fim, quanto à última categoria analítica, a análise dos métodos adotados nos estudos, fez-se necessário desmembrá-la em quatro subcategorias, a saber: ambientes de coleta, procedimentos de coleta (incluindo o uso de instrumentos e *softwares*), procedimentos de análise (incluindo técnicas) e categorias/variáveis bibliométricas mapeadas. Essas subcategorias constituem, na maioria dos casos, elementos adotados em um percurso metodológico de pesquisa.

No que tange aos ambientes de coleta, prevaleceu o uso da Biblioteca Digital Brasileira de Teses e Dissertações (BDTD) do Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT) (10 estudos), seguido pelo banco de dados da CAPES (cinco estudos). Destaca-se que, em nove estudos, foram utilizadas coletas em mais de um ambiente, incluindo: BDTD, banco da CAPES, Portal Domínio Público, repositórios institucionais, dentre outros. O Gráfico 4 apresenta os resultados quanto aos ambientes de coleta dos dados bibliométricos dos estudos.

Gráfico 4 – Ambientes de coleta de dados



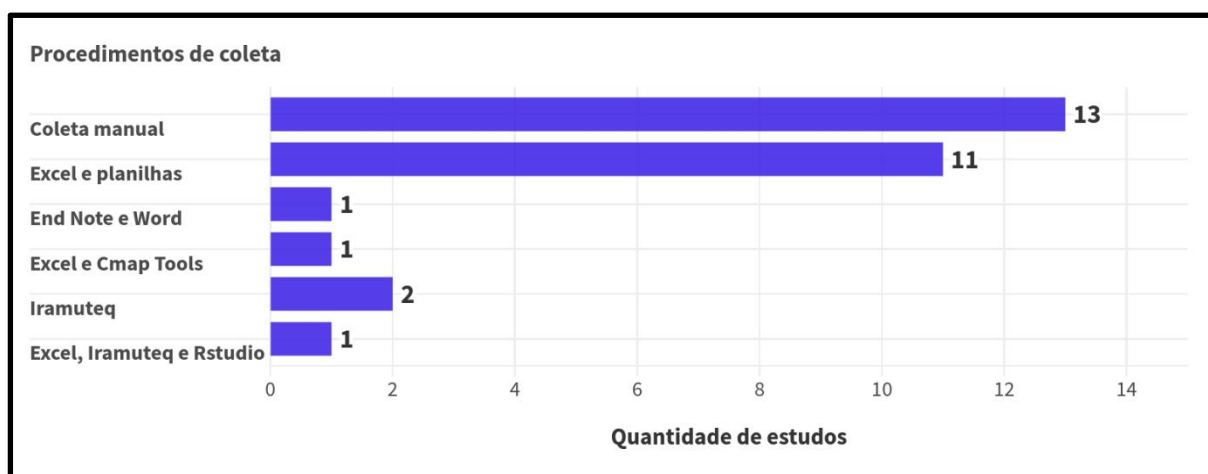
Fonte: elaborado pelo autor (2022).

Os resultados apresentados no Gráfico 4 evidenciam o potencial da BDTD do IBICT e do banco da CAPES quanto ao armazenamento da produção científica dos PPG, em específico o depósito das teses e dissertações defendidas. Segundo Silva, Kronemberger e Gomes (2021, p. 100152), a BDTD é uma base especializada no armazenamento das teses e

dissertações, além de seu potencial tecnológico, pois constitui “[...] um sistema integrado de informação de teses e dissertações existentes nas instituições de ensino e pesquisa brasileiras [...]”. O banco de dados da CAPES, por sua vez, também armazena teses e dissertações, disponibilizando dados sistematizados que facilitam a execução de buscas abrangentes, por conseguinte, promovem a compreensão quanto à origem das pesquisas e de seus desdobramentos (Souza *et al.*, 2013).

Os procedimentos de coleta adotados nos 29 estudos desta revisão são semelhantes, no que tange ao modo de organizar os registros coletados. A maioria das coletas ocorreu de modo manual (13 estudos), ou seja, sem uso de *softwares*, programas ou quaisquer ferramentas tecnológicas para exportação dos dados. Destaque é atribuído ao programa *Excel*, pois foi utilizado em 11 estudos, com a intenção de registrar os dados coletados, mediante a geração de uma planilha de armazenamento e organização. Constata-se o uso de *softwares* especializados para exportação dos dados em, apenas, cinco estudos, conforme ilustrado no Gráfico 5.

Gráfico 5 – Procedimentos de coleta de dados nos estudos da amostra



Fonte: elaborado pelo autor (2022).

Embora a coleta manual predomine na amostra, o uso de *softwares* especializados para coleta e registro de dados, nos levantamentos bibliométricos, tem crescido, provavelmente devido à influência do aperfeiçoamento dessas tecnologias. Essa afirmação foi constatada no estudo de Pereira *et al.* (2019), ao enaltecer o uso do *Excel* como um programa específico para registro dos dados em planilhas, além do uso de *softwares* que possibilitam o processo de migração dos dados, ou seja, com o *software* torna-se possível “[...] transferir os dados da base de dados para que depois possam ser processados e elaboradas as apreciações [...]” (Pereira *et al.*, 2019, p. 11).

Observa-se que a coleta manual foi utilizada em estudos mais antigos da amostra. Isso possibilita afirmar que os estudos mais recentes recorrem a ferramentas informatizadas, de modo a agilizar os procedimentos de pesquisa. Assim, de modo informatizado, é possível ampliar as relações entre os elementos analisados, como: relações de artigos para referências, relações entre palavras-chave, de cocitação entre os autores, de cocitação entre os países, relacionamento de coautoria, dentre outras possibilidades (Pereira *et al.*, 2019). Além disso, com o desenvolvimento tecnológico, é cada vez mais comum combinar as técnicas de coleta, como também utilizar mais de uma ferramenta, com a finalidade de ampliar as análises, executando-as com mais rapidez e precisão, como foi realizado por três estudos da amostra. Portanto, “[...] combinar vários métodos para explorar trajetórias de pesquisa em uma determinada disciplina tornou-se uma tendência na bibliometria [...]” (Pereira *et al.*, 2019, p. 11).

As últimas subcategorias relacionadas aos métodos e técnicas da amostra são os procedimentos de análise e as categorias ou variáveis bibliométricas mapeadas. No conjunto de artigos, revela-se o uso predominante das técnicas matemáticas da estatística descritiva (21 estudos), o que permitiu processar os dados, a fim de gerar resultados, apresentados por meio de estratégias de visualização, conforme já descrito no início desta quinta etapa da revisão integrativa. A técnica da análise de conteúdo foi utilizada em quatro estudos; a mineração dos textos por meio do *software Vantage Point*, em dois estudos; e dois estudos recorreram à combinação de técnicas (um combinou a estatística descritiva com o mapeamento de conceitos e o outro a estatística descritiva com a análise do discurso). As técnicas foram utilizadas sobre categorias ou variáveis específicas, conforme mencionado no Quadro 21.

Quadro 21 – Técnicas de análise de dados e variáveis bibliométricas investigadas

Técnicas de análise de dados	Principais variáveis bibliométricas
Estatística descritiva	Ano, instituição, região geográfica, nível de ensino, área do conhecimento, orientação, coorientação, produtividade autoral, referencial teórico e metodológico, sujeitos da pesquisa, origem do texto, tipologia documental, editora, periódico, regiões brasileiras, discentes, Programa de Pós-Graduação, ano de defesas, produções por níveis de ensino, títulos das publicações, método adotado, temáticas, autores citados, estratégias de coleta e de análise de dados, quantidade de pesquisas, locus da pesquisa, quantidade de orientações e similaridade entre palavras-chave.
Estatística descritiva e análise do discurso	Quantidade de estudos, região geográfica, gênero dos pesquisadores, tipo de estudo, instrumento utilizado, enfoque da pesquisa e aplicação.
Estatística descritiva e mapeamento de conceitos	Quantidade de estudos, temáticas, tipologia documental, ano, instituições, sujeitos de pesquisa, métodos adotados e locus da pesquisa.

Técnicas de análise de dados	Principais variáveis bibliométricas
Mineração de texto (Vantage Point)	Anos, área do conhecimento, Programa de Pós-Graduação, instituição, regiões brasileiras, gênero dos pesquisadores, agências de fomento, temáticas, evolução temporal e orientadores.
Análise de conteúdo	Redes de colaboração entre autores, redes entre instituições, entre Programas de Pós-Graduação, entre tipos de pesquisa, frequência de palavras, objetivo da pesquisa, metodologia, produto final, trabalhos futuros, quantidade de estudos, temáticas, disciplinas do conhecimento, evolução temporal, grandes áreas do conhecimento, subáreas do conhecimento, categorias temáticas, áreas do Programa de Pós-Graduação, temáticas pesquisadas e métodos de coleta e de análise de dados.

Fonte: Dados da pesquisa (2022).

Pelo que consta no Quadro 21, é possível confirmar a diversidade de estratégias que podem ser utilizadas quando se pretende quantificar, avaliar e/ou medir a produção científica de uma área. As técnicas de análise são variadas e, assim como as técnicas de coleta, vêm se aperfeiçoando com o uso de *softwares* especializados, como também com a combinação de técnicas. Acerca desse resultado, dialoga-se com Moraes e Kafure (2020), quando os autores afirmaram que técnicas combinadas e informatizadas potencializam as análises, ampliando as redes de relacionamento. Com efeito, esse processo “[...] facilita a combinação de elementos para entendimento de informações e conhecimentos da área estudada. Possibilita uma visão geral da área, bem como, a exploração de contextos específicos [...]” (Moraes; Kafure, 2020, p. 1).

Embora os resultados do Quadro 21 demonstrem a predominância de técnicas tradicionais de análise, como a estatística descritiva e a análise de conteúdo, o uso do Vantage Point, por exemplo, vem se fortalecendo, pois essa ferramenta é direcionada à mineração, análise e visualização de grandes volumes de dados, podendo ser utilizada de modo combinado com outras técnicas e instrumentos, como os tesauros e os programas para geração de planilhas (Galdino; Garcia; Amaral, 2022). Outro *software* com múltiplas funcionalidades e que cujo uso vem crescendo nas análises bibliométricas é o *VOSviewer*, que pode ser utilizado para exportação de dados das bases (técnicas de coleta), como para gerar redes de relacionamento, além de oportunizar o processo de mineração de texto, condensando, sobretudo, dados qualitativos, mediante a constituição das redes de coocorrência entre as palavras-chave (Galvão; Ferreira; Andrade, 2020).

Com o término da última etapa da revisão, torna-se possível verificar os resultados alcançados e correlacioná-los com a questão de pesquisa formulada no início da revisão. Indagou-se, na questão, sobre os estudos que realizaram mapeamento da produção científica a partir de conteúdos de teses e dissertações. Pelos resultados, a conclusão é de que uma

amostra de 29 estudos foi identificada, contendo estudos com temáticas variadas, perpassando diferentes áreas do conhecimento e com uso diferenciado de metodologias, o que confirma o aspecto interdisciplinar dos estudos bibliométricos e o seu uso recorrente nas pesquisas desenvolvidas nos PPG brasileiros.

Quanto às características dos 29 estudos, os resultados podem ser resumidos a partir das variáveis ou categorias bibliométricas investigadas, sendo revelado o que os pesquisadores estão investigando, por conseguinte, indicam-se possíveis lacunas ainda não resolvidas e que demandam a realização de novos estudos. O Quadro 22 descreve o resumo dos resultados referentes às variáveis bibliométricas que caracterizam a amostra e as proposições para estudos posteriores.

Quadro 22 – Resultados por variáveis bibliométricas e proposições para estudos futuros

Variável	Resultado	Proposições futura
Autorias	Predomínio da autoria coletiva em comparação com a individual.	Incentivar estudos sobre coautoria, considerando suas vantagens na construção colaborativa do conhecimento.
Anos de publicação	Tendência de crescimento do número de publicações ao longo do tempo.	Ampliar os estudos, dado o aumento contínuo da produção bibliométrica e sua relevância.
Visualização de dados	Quadros, tabelas e gráficos prevalecem como os recursos mais utilizados.	Explorar novas formas de visualização, como mapas conceituais e nuvens de palavras.
Redes de colaboração	Poucos estudos realizaram mapeamentos de redes; destacam-se redes de autores, grupos de pesquisa, instituições e domínios.	Ampliar o mapeamento de redes em Programas de Pós-Graduação específicos.
Temáticas	Predomínio de estudos nas áreas da Saúde e Ciências Sociais Aplicadas; na Ciência da Informação, destacam-se Bibliometria, Grupos de Pesquisa, Comunicação Científica e Taxonomias.	Investir em novas abordagens nessas áreas e ampliar investigações em Engenharias e Ciências Exatas; na Ciência da Informação, mapear a produção dos PPG da área.
Objetivos e Conclusões	Fundamentam-se na trindade bibliométrica: quantificar, medir e avaliar a produção do conhecimento.	Reforçar os fundamentos bibliométricos como base contínua de investigação.
Métodos adotados	A BDTD e a base da CAPES são as fontes mais utilizadas; Excel e planilhas são as principais ferramentas de coleta; prevalência da estatística descritiva e da análise de conteúdo.	Manter o uso da BDTD e CAPES, diversificando ferramentas (como Cmap Tools, Iramuteq, RStudio); combinar técnicas com mapeamento de conceitos, mineração de textos e <i>softwares</i> como VantagePoint e VOSviewer.

Fonte: Dados da pesquisa (2022).

A partir dos resultados alcançados e da identificação das lacunas constantes no Quadro 22, observa-se um campo fértil para o desenvolvimento de estudos inéditos, especialmente aqueles que combinem abordagens bibliométricas com estratégias de tratamento sistemático de grandes volumes de dados. Essas lacunas fundamentam a pertinência da presente investigação, que, ao se articular com os desafios identificados, busca contribuir para o avanço das análises sobre a produção científica da pós-graduação brasileira, com foco na Ciência da Informação.

4 METODOLOGIA

Neste capítulo são apresentados os métodos, as técnicas e os recursos utilizados para a responder aos objetivos da pesquisa. No contexto científico ou acadêmico, a metodologia consiste no conjunto de procedimentos aplicados para alcance de resultados pautados em evidências científicas. Assim, ela “desenvolve-se ao longo de um processo que envolve inúmeras fases, desde a adequada formulação do problema até a satisfatória apresentação dos resultados” (Gil, 2002, p. 17). Diehl e Tatim (2004) comparam a metodologia científica a uma bússola, no entendimento de que, durante as fases que formam o processo, serão estudados e avaliados os métodos a serem adotados, para que o pesquisador possa identificar as possibilidades e limitações dentro de um escopo de análise. A partir disso, entende-se que os procedimentos metodológicos indicam uma melhor direção dentro do período da pesquisa.

4.1 Caracterização da pesquisa

Caracteriza-se esta pesquisa como aplicada, cuja finalidade não se limita a aspectos intelectuais e/ou reflexivos, mas contempla uma concepção empírica (pautada na experiência ou no levantamento de evidências práticas). Segundo Diehl e Tatim (2004), ao contrário da pesquisa pura, que busca pelo progresso científico a partir do conhecimento teórico, a pesquisa aplicada caracteriza-se pelo interesse prático, cujos resultados poderão ser aplicados para solução de problemas que permeiam o cotidiano das pessoas, das organizações e da sociedade, em geral.

De maneira mais particular, segundo os autores citados, é possível caracterizar a pesquisa do ponto de vista dos seus objetivos. Sobre esse tipo de classificação, Gil (2002) defende a existência de pesquisas exploratórias, descritivas e explicativas. Para este estudo, considera-se que ele se classifica como exploratório e descritivo. Justifica-se essa classificação, pois, as pesquisas exploratórias, segundo Gil (2002), visam ao levantamento exaustivo de outras investigações, com o propósito de situar o investigador no universo da temática. Este estudo também se apresenta como descritivo, porque apresenta as características do objeto analisado e estabelece relações entre variáveis, a fim de facilitar a compreensão do objeto (Diehl; Tatim, 2004).

Quanto aos procedimentos técnicos, seguindo a classificação proposta por Gil (2002), esta investigação recorre ao estudo bibliométrico, do ponto de vista prático ou procedimental. Os estudos bibliométricos recorrem a técnicas estatísticas “[...] de mediação de objetos de

estudos complexos, quantificando e descrevendo, de modo a favorecer prognósticos acerca dos processos de investigação e comunicação escrita de um determinado tema [...]” ou área do conhecimento (Brilhante *et al.*, 2016, p. 705). Também utiliza a pesquisa documental para a coleta de dados na base de dados abertos da CAPES, referente à produção dos PPG brasileira, com foco na área da Ciência da Informação. Na pesquisa documental, trabalha-se com dados da CAPES que, segundo Sá-Silva *et al.*, (2009), são dados secundários. Isso se justifica, conforme a visão que o autor defende, porque a CAPES disponibiliza dados que já foram processados e agregados a partir de outras fontes (como informações fornecidas por universidades e programas de pós-graduação), incluindo relatórios consolidados sobre a produção científica, indicadores de desempenho de programas, dados extraídos de repositórios institucionais.

Quanto à abordagem, os estudos podem ser classificados em quantitativos, cujos dados são mensurados, a partir de regras estatísticas e considerando a objetividade e exatidão numérica, e em qualitativos, com análises subjetivas, que vão além da quantificação (Creswell, 2014). Nesta tese, pretende-se coletar dados quantitativos, que serão agrupados em categorias ou variáveis capazes de representar o panorama científico dos PPG brasileiros da área de Ciência da Informação, utilizando análises de natureza prioritariamente descritiva, centrado na mensuração, tabulação e visualização dos dados coletados. Nesta pesquisa ocorre na análise exploratória dos dados empíricos, permitindo identificar padrões emergentes, tendências e possíveis particularidades da pós-graduação em Ciência da Informação dentro do contexto nacional.

A pesquisa adota um certo aspecto comparativo, estabelecendo relações entre o panorama geral da pós-graduação *stricto sensu* no Brasil e a realidade dos programas da Ciência da Informação. A comparação é conduzida de forma analítica, sem o uso de técnicas estatísticas específicas, priorizando a interpretação dos indicadores quantitativos. Os dados analisados incluem o número de programas e titulados, distribuição regional, tempo médio de titulação e evolução histórica, assim como a produção. A partir dessa abordagem, busca-se não apenas descrever a posição da Ciência da Informação na pós-graduação nacional, mas também identificar possíveis implicações para políticas acadêmicas e tendências futuras.

4.2 Os dados abertos Capes

No Brasil, a política de dados abertos foi instituída pelo Decreto nº 8.777, de 11 de maio de 2016 (Brasil, 2016), que, em seu Art. 2º, inciso III, define dados abertos como “dados

acessíveis ao público, representados em meio digital, estruturados em formato aberto, processáveis por máquina, referenciados na internet e disponibilizados sob licença aberta”, e, com isso, que seja possível a “sua livre utilização, consumo ou cruzamento, limitando-se a creditar a autoria ou a fonte”. Na CAPES, para atender ao disposto no referido Decreto, foram elaborados PDA que compilam normas, ações e estratégias sistematizadas da instituição, que norteiam as atividades de implementação, desenvolvimento, manutenção e promoção da abertura de dados, sendo que o último PDA publicado refere-se ao biênio dezembro de 2022 a dezembro de 2024. A priorização dos conjuntos de dados a serem disponibilizados foi definida considerando a relevância das informações para o cidadão, o grau de maturidade dos conjuntos de dados dentro da instituição e a Lei Geral de Proteção dos Dados (LGPD) (Capes, 2022b).

A LGPD é uma legislação brasileira que estabelece regras para o tratamento de dados pessoais no país. Ela entrou em vigor em 18 de setembro de 2020 e tem como objetivo proteger os direitos fundamentais de liberdade e privacidade dos indivíduos, estabelecendo princípios e obrigações para as empresas e órgãos públicos que tratam dados pessoais (Frazão; Oliva; Tepedino, 2019).

A LGPD se aplica a qualquer tipo de dado pessoal, incluindo dados abertos. Ela não impede a disponibilização de dados abertos, mas estabelece regras para garantir a privacidade e a segurança dos dados pessoais, mesmo quando esses dados são disponibilizados publicamente. Isso inclui regras para o armazenamento, o compartilhamento e o uso dos dados, bem como regras para a transparência e a responsabilidade dos órgãos e empresas que tratam esses dados (Gregori, 2020). É importante que as organizações e órgãos públicos que disponibilizam dados abertos estejam cientes das obrigações estabelecidas pela LGPD e as cumpram para garantir a proteção dos dados pessoais (Cristóvam; Hahn, 2020).

Dentro deste contexto, o tratamento de dados pessoais se tornou uma obrigação dos responsáveis pela sua guarda, pois esses dados costumam ser usados para fins científicos, comerciais, políticos, entre outros. Logo, as informações devem ser tratadas de forma segura, evitando-se o vazamento de dados, a invasão de computadores e outras ameaças à privacidade e à segurança dos usuários (Souza *et al.*, 2018). Para isso, existem algumas medidas que podem ser adotadas pelos responsáveis pelo tratamento de dados pessoais, tais como: criptografia de dados, controle de acesso, auditorias de segurança, formação dos usuários, políticas de segurança, backup dos dados e monitoramento do ambiente (Piurcosky *et al.*, 2019). Assim, o tratamento de dados pessoais é fundamental para garantir a privacidade dos usuários e sua segurança, seguindo as melhores práticas de proteção de dados.

Os usuários do portal Dados Abertos CAPES podem acessar informações sobre Acessos ao Portal de Periódicos, Bolsas e Auxílios, Avaliação da Pós-Graduação *stricto sensu* e Servidores, Contratos, Dotação e Execução Orçamentária da Capes. Os dados disponíveis são publicados em formato aberto e podem ser baixados e usados para diversos fins. Esses dados representam o universo desta pesquisa.

Considerando as disposições do PDA, a CAPES disponibiliza, em seu sítio eletrônico, 57 conjuntos de dados. Dentre esses conjuntos de dados há dados e informações sobre a pós-graduação brasileira, que é o objeto deste estudo. Esse conjunto apresenta subconjuntos de dados referentes à avaliação da Pós-Graduação *stricto sensu*, sendo que foi identificada a existência de 32 *datasets*.

Os dados disponibilizados pelo portal Dados Abertos CAPES, como o Conjunto de Dados referente aos “Dados de Avaliação dos Programas de Pós-Graduação *stricto sensu*”, permitem a análise da performance dos programas de pós-graduação *stricto sensu*. Os autores Torino, Trevisan e Vidotti (2019) realizaram uma análise desses dados para discutir o atendimento de 29 conjuntos de dados abertos governamentais acerca da Avaliação da Pós-Graduação *stricto sensu* disponível no portal Dados Abertos CAPES, aos 13 desafios para a publicação de dados na web recomendados pelo W3C.

Torino, Trevisan e Vidotti (2019) argumentam que os dados disponibilizados no portal não estão de acordo com as cinco estrelas dos dados abertos, estabelecidos por Berners-Lee (2012), mas que eles atendem à terceira estrela, onde esta exige que os dados sejam disponibilizados sob uma licença aberta, permitindo que qualquer pessoa os utilize, compartilhe e combine livremente.

4.2.1 Dados Abertos Governamentais (DAG)

A disponibilização de Dados Abertos Governamentais (DAG) tem sido uma diretriz importante para ampliar a transparência e democratizar o acesso à informação, permitindo que pesquisadores, jornalistas e cidadãos utilizem esses dados para diversas finalidades (Ubaldo, 2013). No Brasil, a Política Nacional de Dados Abertos estabelece que órgãos e entidades da administração pública devem divulgar seus conjuntos de dados em formato aberto, promovendo a transparência e incentivando a inovação (Brasil, 2016). No contexto da educação e da ciência, órgãos como a CAPES disponibilizam informações detalhadas sobre programas de pós-graduação, docentes e discentes, permitindo análises acadêmicas e bibliométricas baseadas em dados oficiais (Capes, 2024).

Esta pesquisa utiliza os dados da CAPES como fonte principal para a análise bibliométrica, seguindo os princípios da Ciência Aberta. Entretanto, não foram objeto deste estudo discussões sobre a qualidade, interoperabilidade ou desafios da reutilização desses dados, que foram discutidos em Brito *et al.* (2015). Embora a disponibilidade de informações estruturadas seja um avanço, a literatura ressalta que a mera abertura dos dados não garante sua efetiva reutilização, pois aspectos como padronização, metadados descritivos e mecanismos de atualização são essenciais para garantir sua aplicabilidade em diferentes contextos (Brito *et al.*, 2015). Nesse sentido, os princípios *FAIR* (*Findable, Accessible, Interoperable, and Reusable*) surgem como um referencial para a disponibilização de dados científicos de forma acessível e reutilizável (Wilkinson *et al.*, 2016), sendo um fator crítico para estudos bibliométricos.

No caso específico da CAPES, os dados fornecidos são utilizados para análises e estudos acadêmicos, mas apresentam desafios relacionados à estruturação e padronização, conforme destacado no capítulo do referencial teórico. A falta de padronização pode impactar a reprodutibilidade de pesquisas e a comparação entre diferentes períodos ou instituições. Neste estudo, os dados foram tratados e organizados para permitir análises bibliométricas, mas não foram abordadas questões relativas à governança dos dados, sua curadoria ou políticas de preservação e atualização.

A relação entre dados abertos e estudos bibliométricos é um tema relevante, pois a qualidade e a organização das bases de dados influenciam diretamente a confiabilidade dos resultados obtidos. A bibliometria depende de informações precisas e estruturadas, e a ausência de padrões nos dados abertos pode limitar as análises e comprometer inferências sobre produtividade e colaboração científica (Brito *et al.*, 2015). Assim, a presente pesquisa beneficia-se dos dados disponibilizados pela CAPES, mas reconhece a necessidade de investigações futuras que analisem a estrutura e a qualidade desses dados, contribuindo para aprimorar sua utilização em estudos científicos.

4.3 Delimitação do objeto da pesquisa

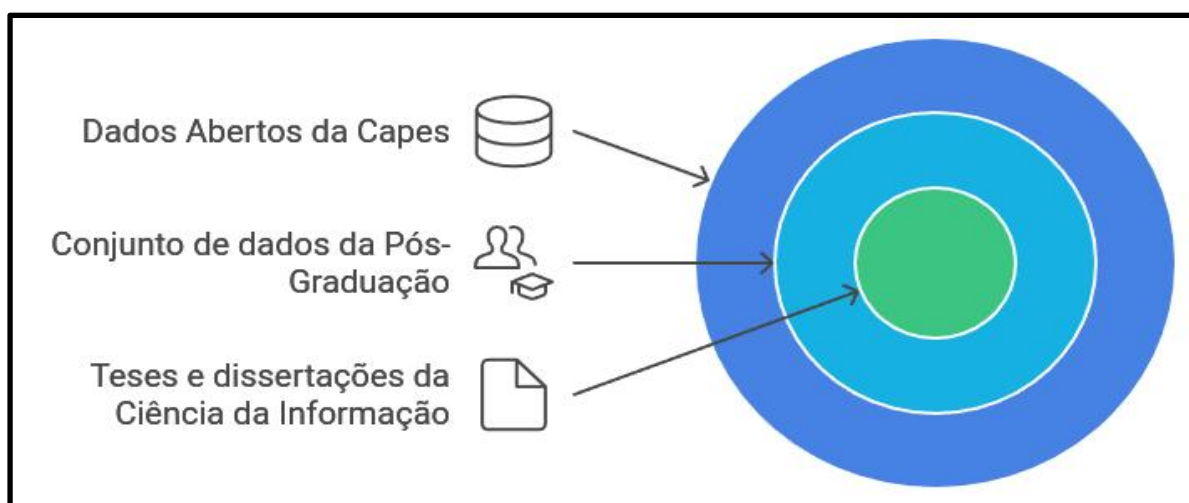
Para a coleta de dados para a realização desta pesquisa foi selecionado o Catálogo de Teses e Dissertações (CTD) da CAPES, instituído no âmbito do PDA da CAPES. A delimitação do objeto constitui o recorte temático e contextual, com a finalidade de tornar o objeto de estudo mais específico, como também facilitar os procedimentos de coleta e de

análise de dados. De acordo com Marconi e Lakatos (2021), a delimitação permite tratar o tema em profundidade e abordar diferentes possibilidades de compreensão do objeto.

Considerando que a Ciência da Informação está em desenvolvimento no Brasil (Queiroz; Vilan Filho; Moura, 2015), e que adota métodos científicos para melhorias quanto ao tratamento da informação e uso adequado (Saracevic, 1996, p. 43), sendo “uma atividade essencialmente aplicada” (Silva; Ribeiro, 2020, p. 55), delimita-se, desse escopo abrangente (Ciência da Informação brasileira), uma amostra representativa, os PPG dessa área, com foco nos aspectos que identificam esses programas e as produções deles por meio da defesa das teses e dissertações.

A Figura 3 apresenta a delimitação do objeto e contexto analisado nesta tese (universo e amostra).

Figura 3 – Delimitação do objeto estudado: universo, objeto e delimitação



Fonte: Dados da pesquisa (2022).

Conforme apresentado na Figura 3, o universo desta pesquisa está representado pelos dados abertos disponibilizados pela Capes em 57 conjuntos de dados. O objeto da pesquisa são os dados abertos sobre a Avaliação da Pós-Graduação *stricto sensu* brasileira, e a delimitação do objeto se dá pela análise dos dados de avaliação sobre a Ciência da Informação brasileira, entre os anos de 2004 a 2022, composta por um conjunto de elementos, tais como: os cursos de Graduação e Programas de Pós-Graduação vinculados à área, os periódicos científicos, os eventos, as associações profissionais e científicas, dentre outros. Dentre esses elementos, delimitou-se como recorte os PPG, que serão mapeados, tendo suas principais características identificadas como: ano de defesa, as instituições, número de estudos defendidos, informações dos docentes e discentes. No âmbito desses programas, serão

analisadas as teses e dissertações defendidas, com foco nas temáticas, autorias, colaboração nas orientações e nas bancas de defesa, dentre outras variáveis relativas a essas produções. Portanto, os PPG e os estudos neles defendidos constituem o recorte e delimitação de análise. Os componentes desse conjunto de teses e dissertações foram recuperados em um ambiente público que disponibiliza dados abertos a respeito da Pós-Graduação brasileira, que é o CTD da CAPES.

Dos 32 *datasets* identificados nos subconjuntos de dados disponibilizados pela Capes, referentes à avaliação da Pós-Graduação *stricto sensu*, foram escolhidos os quatro CTD que contêm dados referentes aos anos de 1987 a 2022.

O primeiro CTD refere-se às teses e dissertações defendidas entre 1987 e 2022, onde são disponibilizados dados ou variáveis referentes a: resumo, palavras-chave, nome do discente, nome da instituição de ensino, unidade federativa, entre outros. O segundo CTD abrange os trabalhos de 2013 a 2016 e contém 56 tipos de dados, sendo adicionados códigos de identificação para instituição, discente, docente, data de entrada e saída do programa, o endereço para o texto completo da dissertação ou tese na Plataforma Sucupira, dentre outros detalhes. No terceiro CTD, os dados referem-se aos anos de 2017 a 2020 e possuem 58 variáveis. Por fim, o quarto CTD se refere aos anos de 2021 a 2024, abrangendo o mesmo número de variáveis do anterior, contendo disponíveis até o momento apenas os dados para o ano de 2022. Neste catálogo, foram adicionados mais dois campos em relação ao segundo, que mostram a existência de estudos vinculados a essa produção (como artigos de periódicos, de eventos, dentre outros) e, caso haja, qual é a identificação em outra base de dados (Capes, 2022a).

Como se trata de uma pesquisa documental e bibliométrica, cuja amostra de dados será gerada pela extração ou levantamento de dados coletados e disponibilizados pela CAPES, que estão abertos e disponibilizados em plataformas eletrônicas (dados disponibilizados no CTD), é conveniente o uso de ferramentas informatizadas para apoiar o trabalho do pesquisador, como recomendado por Marconi e Lakatos (2021). A seção seguinte aborda os insumos tecnológicos utilizados durante a condução do estudo.

4.4 Insumos tecnológicos

A parte empírica da pesquisa, sobretudo no que tange à coleta de dados (extração dos dados do CTD), requer o uso de *softwares* e linguagens de programação. Para esta parte da pesquisa foram utilizados os seguintes insumos tecnológicos:

a) O OpenRefine⁸ (versão 3.6.1), nas versões anteriores denominado de *GoogleRefine*, é um *software* gratuito e de código aberto, escrito em *Java*, compatível com qualquer sistema operacional (por meio do navegador), lançado em 2010, que ajuda os usuários a gerenciar os conjuntos de dados dispersos e também a limpar as inconsistências que eventualmente levarão a erros, fornecendo os meios para importar conjuntos de dados (Kusumasari *et al.*, 2016). Assim, ele é usado para transformar dados confusos em dados limpos e utilizáveis (Verborgh; De Wilde, 2013). O *software* funciona como um servidor *web* no computador do usuário e utiliza o navegador de internet como interface, mas os dados permanecem locais. Trata-se, assim, de um programa utilitário poderoso, capaz de transformar e limpar grandes conjuntos de dados. No entanto, o *OpenRefine* pode manifestar-se como uma ferramenta complexa, munida de linguagem de programação e expressões computacionais, o que exigirá, em especial dos usuários sem experiência, um elevado nível de aprendizado computacional (Ham, 2013).

b) O *Calc* do *LibreOffice* (versão 7.4.1.2) é um *free software*(software livre) que faz parte do conjunto de editores de texto do pacote *LibreOffice*, cujo propósito principal é oportunizar a criação, edição e apresentação de planilhas eletrônicas (Afonso *et al.*, 2014). Ele permite a criação, edição, análise e visualização de dados tabulares, sendo uma alternativa gratuita ao *Microsoft Excel*. Compatível com formatos como *OpenDocument Spreadsheet* (ODS) e *XLSX* (*Excel*), o *Calc* possibilita a utilização de fórmulas, gráficos, tabelas dinâmicas e macros para automação de tarefas. Entre seus recursos, destaca-se a inserção e manipulação de dados numéricos, a utilização de funções e fórmulas para cálculos matemáticos e estatísticos, a organização e filtragem de dados por meio de funções de banco de dados, além do uso de ferramentas estatísticas para análise de informações. O *software* também possibilita a criação de gráficos dinâmicos, a automação de tarefas repetitivas por meio de macros e a ocultação de colunas ou linhas sem exclusão permanente. No que se refere à importação e exportação de planilhas, o *Calc* permite modificar arquivos do *Microsoft Excel*, integrar bases de dados empresariais, incorporar fluxos de dados em tempo real em relatórios e exportar planilhas no formato PDF (Machado, 2024).

c) Ferramentas e técnicas da Ciência de Dados: a pesquisa utilizou ferramentas e técnicas da Ciência de Dados para tratar, analisar e visualizar os dados coletados. Esses insumos tecnológicos viabilizaram a manipulação de grandes volumes de informação provenientes dos conjuntos de dados analisados. Entre as técnicas empregadas, destacam-se:

⁸ Disponível em: <https://openrefine.org/documentation.html>. Acesso em: 3 novembro 2022.

Mineração de Dados: para extrair informações relevantes a partir dos registros analisados, permitindo a identificação de padrões e tendências dentro do conjunto de dados; NLP: para a limpeza, organização e análise textual de dados provenientes de documentos e metadados, otimizando a classificação e a extração de termos representativos; Estatística Computacional: para a análise quantitativa dos dados, incluindo distribuição de frequências, medidas de tendência central e dispersão, fundamentais para a interpretação dos resultados; e Ferramentas de Visualização de Dados: para apresentar os achados da pesquisa de forma intuitiva, facilitando a interpretação dos resultados e permitindo a exploração de relações entre variáveis (Moraes; Kafure, 2020).

A manipulação e importação dos dados coletados são realizadas utilizando a linguagem de programação *Python* (versão 3.x), devido à sua adoção em Ciência de Dados e suporte a múltiplos paradigmas de programação (Vanrossum; Drake, 2010; Biswas; Chand, 2019). Para a preparação, transformação e visualização dos dados, foram empregadas cinco bibliotecas principais:

- *Pandas* (versão 1.4.2): utilizada para manipulação e análise de dados, permitindo a leitura de diferentes formatos e a organização de *DataFrames* (Mckinney, 2012).
- *Seaborn* (versão 0.11.2): aplicada à plotagem de gráficos para análise exploratória de dados (Pazos; Sirqueira, 2022).
- *Matplotlib* (versão 3.5.0): utilizada na criação e visualização de gráficos (Pazos; Sirqueira, 2022).
- *Datetime* (versão 3.10.4): empregada na manipulação de datas e horas (Chen, 2018).

O processamento dos dados foi realizado em um ambiente on-line para garantir segurança e acessibilidade. O *Google Colaboratory*⁹ (*Google Colab PRO*) foi utilizado como ambiente de desenvolvimento, possibilitando a execução de código em *Jupyter Notebooks* diretamente na nuvem, sem necessidade de configuração local. Além de permitir colaboração e compartilhamento de código, o *Google Colab* oferece suporte à instalação de bibliotecas adicionais (Alves; Vieira, 2019; Gunawan *et al.*, 2020). Entretanto, sua versão gratuita apresenta restrições quanto ao uso de memória e processamento (Alves; Vieira, 2019).

⁹ Disponível em: <https://colab.google/>. Acesso em: 4 mar. 2025.

d) Ferramentas para visualização de dados: Para a visualização dos dados, foram utilizadas as plataformas *Flourish Studio*¹⁰ e *Datawrapper*¹¹, ambas empregadas para a criação de representações gráficas e mapas interativos. O *Flourish Studio* é uma ferramenta versátil que permite a elaboração de visualizações dinâmicas, dashboards interativos e gráficos personalizados, sendo adotado em diversas áreas do conhecimento (Priya, 2023). O *Datawrapper* foi utilizado para a construção de mapas, possibilitando a transformação dos dados em representações visuais claras e acessíveis. A escolha dessas ferramentas se deve à sua disponibilidade gratuita e à facilidade de uso, permitindo a conversão eficiente dos dados em gráficos e mapas (Islam; Jin, 2019).

A combinação desses insumos tecnológicos permite estruturar a base de dados e extrair informações que sustentam as análises realizadas, garantindo maior confiabilidade e precisão na interpretação dos resultados.

4.5 Etapas e procedimentos metodológicos

Para facilitar a condução desta pesquisa, são propostas etapas, que contêm a descrição dos procedimentos práticos em cada uma delas, de acordo com os quais os resultados serão apresentados. As etapas metodológicas, de acordo com Marconi e Lakatos (2021), indicam o caminho ou percurso que é seguido para alcance dos objetivos da pesquisa, explicitando, sobretudo, os instrumentos de coleta e de análise de dados.

4.5.1 O estudo bibliométrico

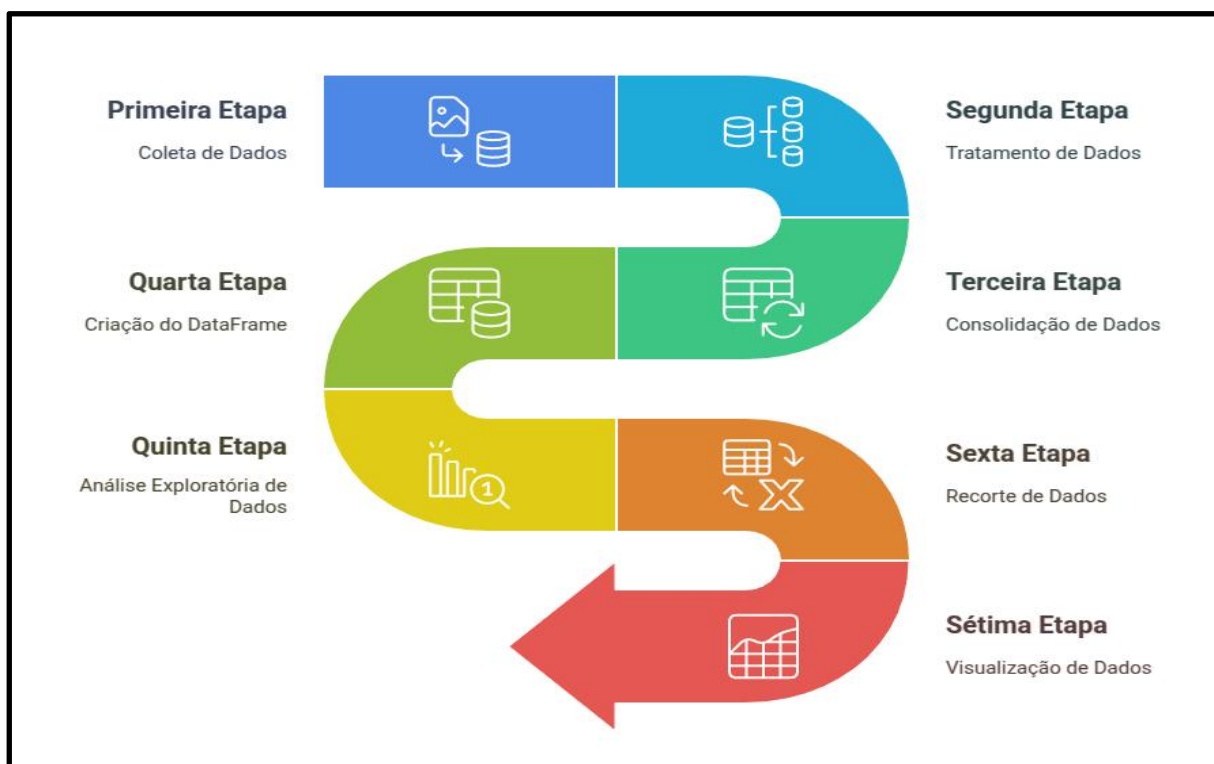
Esta seção descreve os métodos e as técnicas adotadas para conduzir a pesquisa, de modo a garantir a validade e a confiabilidade dos resultados obtidos. A escolha dos procedimentos metodológicos está fundamentada na natureza do problema de pesquisa e nos objetivos previamente estabelecidos, buscando proporcionar uma abordagem sistemática e rigorosa ao tratamento das questões investigadas.

Inicialmente, é apresentado um resumo do delineamento dos procedimentos e dos critérios de seleção de dados, conforme mostra a Figura 4, seguido pelo detalhamento das etapas propostas para o estudo.

¹⁰ Disponível em: <https://app.flourish.studio/>. Acesso em: 4 mar. 2025.

¹¹ Disponível em: <https://app.datawrapper.de/>. Acesso em: 4 mar. 2025.

Figura 4 – Fases da pesquisa bibliométrica



Fonte: Elaborada pelo autor (2025).

A Figura 4 apresenta as etapas do processo de análise dos dados, iniciando pela Coleta de Dados, na qual os arquivos são obtidos no Portal de Dados Abertos da CAPES. Essa etapa contempla a aquisição de informações relacionadas à Pós-Graduação *stricto sensu* no Brasil, abrangendo os conjuntos de dados do Catálogo de Teses e Dissertações, Docentes, Discentes e Programas no período de 2004 a 2022. Em seguida, ocorre a Tratamento dos Dados, que compreende a identificação e correção de valores ausentes, padronização da codificação dos arquivos e harmonização das variáveis para garantir a consistência dos registros.

A terceira etapa corresponde à Consolidação dos Conjuntos de Dados, na qual as diferentes bases são integradas e organizadas para estruturar um conjunto unificado de informações. Após essa fase, procede-se à Criação do DataFrame, onde os dados consolidados são validados quanto à integridade e coerência. A quinta etapa refere-se à Análise Exploratória e Ajustes dos Dados, que envolve a aplicação de técnicas para identificação de padrões e tendências nos registros, bem como ajustes necessários para aprimorar a qualidade das informações. Na sequência, realiza-se o Recorte dos Dados, etapa em que são selecionados apenas os registros pertinentes à Ciência da Informação, delimitando o escopo da investigação. Por fim, os resultados são apresentados por meio da visualização dos dados, permitindo uma compreensão detalhada das características e distribuições observadas nos

conjuntos analisados. O detalhamento de cada uma dessas etapas é apresentado ao longo deste capítulo.

4.5.1.1 Etapa 1 - Coleta de dados

O Portal de Dados Abertos da CAPES foi acessado e foram baixados 19 arquivos do site <https://dadosabertos.capes.gov.br/>. Estes arquivos estão relacionados ao tema "Avaliação da Pós-Graduação Stricto Sensu" e pertencem especificamente aos conjuntos "Catálogo de Teses e Dissertações - Brasil", "Docentes da Pós-Graduação Stricto Sensu no Brasil", "Discentes da Pós-Graduação *stricto sensu* do Brasil" e "Projetos dos Programas de Pós-Graduação Stricto Sensu no Brasil" referentes aos anos compreendidos entre 2004 e 2022. Uma análise detalhada revelou que os conjuntos de dados de 2004 a 2012 contêm menos campos em comparação com os arquivos de 2013 a 2022. Dada esta disparidade, foi preciso adaptar e harmonizar os campos para possibilitar o acoplamento dos DataFrames. No decorrer desta investigação, os conjuntos de dados de 2004 a 2012 foram denominados como 'DF1', enquanto os de 2013 a 2022 foram denominados como 'DF2'. A meta subsequente é acoplar o DF1 ao DF2, resultando, assim, no DataFrame DF3_trabalhos_BR, que compreenderá os registros unificados.

4.5.1.2 Etapa 2 – Tratamento dos dados

O tratamento dos dados envolve um conjunto de operações destinadas a organizar, limpar e preparar os dados coletados para a análise. Os arquivos agrupados sob a denominação DF1 englobam os anos de 2004 a 2012, e estão delineados na Tabela 1.

Tabela 1 – Nome dos arquivos que representam o DF1

Nome do arquivo	Tamanho	Quantidade de linhas
dados_2004	81.6 MB	34917
dados_2005	93.5 MB	39695
dados_2006	98.9 MB	41647
dados_2007	103.4 MB	42819
dados_2008	114.9 MB	46750
dados_2009	124.2 MB	50167
dados_2010	127.8 MB	50903

Nome do arquivo	Tamanho	Quantidade de linhas
dados_2011	140.9 MB	55554
dados_2012	156.9 MB	61053

Fonte: Dados de pesquisa (2024).

Ao analisar a Tabela 1, nota-se a evolução temporal destes conjuntos de dados, é perceptível uma tendência crescente tanto no que se refere ao tamanho dos arquivos quanto à quantidade de registros que eles contêm. Esta última métrica, do número de registros, pode ser interpretada como uma representação direta do número de trabalhos defendidos anualmente, indicando, assim, um aumento da produção acadêmica no período em questão.

O conteúdo interno de DF1 é detalhado e categorizado por meio de 41 campos distintos, conforme ilustrado na Figura 5.

Figura 5 – Nome dos campos encontrados nos arquivos do DF1

```
AnoBase;CodigoPrograma;Regiao;Uf;SiglaIes;NomeIes;
NomePrograma;GrandeAreaCodigo;GrandeAreaDescricao;
AreaConhecimentoCodigo;AreaConhecimento;AreaAvaliacao;
DocumentoDiscente;Autor;TituloTese;Nivel;DataDefesa;
PalavrasChave;Volume;NumeroPaginas;BibliotecaDepositaria;
Idioma;ResumoTese;LinhaPesquisa;URLTextoCompleto;
DocumentoOrientador_1;Orientador_1;
DocumentoOrientador_2;Orientador_2;
DocumentoOrientador_3;Orientador_3;
DocumentoOrientador_4;Orientador_4;
DocumentoCoOrientador_1;CoOrientador_1;
DocumentoCoOrientador_2;CoOrientador_2;
DocumentoCoOrientador_3;CoOrientador_3;
DocumentoCoOrientador_4;CoOrientador_4
```

Fonte: Dados de pesquisa (2024).

Os campos expostos na Figura 5 abrangem uma variedade de informações pertinentes às Teses e Dissertações defendidas na Pós-Graduação entre os anos supracitados (2004-2012). Além disso, é relevante observar que a natureza dos dados varia, compreendendo tanto entradas numéricas quanto alfanuméricas. Tal detalhamento é necessário para uma análise minuciosa e para a identificação de tendências, desafios e oportunidades no contexto da pós-graduação brasileira durante o período em foco.

A Tabela 2 apresenta os conjuntos de dados definidos como DF2, referentes ao período de 2013 a 2022.

Tabela 2 – Nome dos arquivos que representam o DF2

Nome do arquivo	Tamanho	Quantidade de linhas
dados_2013	329.6 MB	67534
dados_2014	339.7 MB	70668
dados_2015	369.2 MB	75884
dados_2016	410.4 MB	80278
dados_2017	449.0 MB	85310
dados_2018	479.5 MB	90469
dados_2019	504.5 MB	94503
dados_2020	434.8 MB	80114
dados_2021	451.5 MB	82240
dados_2022	453.3 MB	82367

Fonte: Dados de pesquisa (2024).

A Tabela 2 apresenta os conjuntos de dados definidos como DF2, referentes ao período de 2013 a 2022. Semelhantemente ao observado na Tabela 1, há uma tendência de crescimento tanto no tamanho dos arquivos quanto na quantidade de registros anuais. No entanto, merece destaque o declínio evidente em 2020, em ambas as métricas, acompanhado de um leve aumento nos anos de 2021 e 2022, mas não atinge o nível de 2019. Uma hipótese plausível para esta redução é o impacto da pandemia da Covid-19, caracterizada pela Organização Mundial da Saúde (OMS) em 11 de março de 2020, que repercutiu na produção acadêmica e na apresentação e defesa de trabalhos no período (Assunção-Luiz, 2021). Interessantemente, apesar de uma ligeira recuperação em 2021 e 2022, os dados de 2022, com 82.367 registros, indicam que a pós-graduação ainda não havia retomado completamente seu ímpeto pré-pandemia. Acredita-se que, durante a pandemia, muitos estudantes enfrentaram dificuldades para conduzir suas pesquisas, especialmente aquelas que dependiam de coleta de dados presencial, acesso a laboratórios, bibliotecas ou trabalho de campo. Esses atrasos em 2020 e 2021 possivelmente tenha estendido o tempo necessário para concluir teses e dissertações, resultando em um número menor de defesas também em 2022.

A estrutura interna dos arquivos de DF2 é retratada na Figura 6, onde estão elencados os nomes das colunas. Contrapondo-se aos arquivos do DF1, os DataFrames de DF2 têm número de campos diferenciados, pois, para o intervalo de 2013 a 2016, são compostos por 56 campos - uma expansão de 15 campos em relação ao DF1, e para os registros de 2017 a 2022, o número de campos ascende a 58, marcando a adição de dois novos campos.

Figura 6 – Nome dos campos encontrados nos arquivos do DF2

```

AN_BASE; CD_PROGRAMA; NM_PROGRAMA; SG_ENTIDADE_ENSINO;
NM_ENTIDADE_ENSINO; ID_ADD_PRODUCAO_INTELCTUAL;
ID_PRODUCAO_INTELCTUAL; NM_PRODUCAO; ID_SUBTIPO_PRODUCAO;
NM_SUBTIPO_PRODUCAO; ID_AREA_CONCENTRACAO; NM_AREA_CONCENTRACAO;
ID_LINHA_PESQUISA; NM_LINHA_PESQUISA; ID_PROJETO; NM_PROJETO;
DH_INICIO_AREA_CONC; DH_FIM_AREA_CONC; DH_INICIO_LINHA;
DH_FIM_LINHA; DT_TITULACAO; DS_PALAVRA_CHAVE; DS_ABSTRACT;
DS_KEYWORD; IN_TRABALHO_MESMA_AREA; NM_TP_VINCULO;
IN_ORIENT_PARTICIPOU_BANCA; DS_BIBLIOTECA_DEPOSITARIA;
ID_TP_EXPECTATIVA_ATUACAO; NM_EXPECTATIVA_ATUACAO;
ID_PESSOA_DISCENTE; NM_DISCENTE; DT_MATRICULA;
ID_GRAU_ACADEMICO; NM_GRAU_ACADEMICO; NM_ORIENTADOR;
DS_CATEGORIA_ORIENTADOR; NM_CATEGORIA_DOCENTE; NM_REGIAO;
SG_UF_IES; NM_UF_IES; CD_GRADE_AREA_CONHECIMENTO;
NM_GRADE_AREA_CONHECIMENTO; CD_AREA_CONHECIMENTO;
NM_AREA_CONHECIMENTO; CD_SUBAREA_CONHECIMENTO;
NM_SUBAREA_CONHECIMENTO; CD_ESPECIALIDADE; NM_ESPECIALIDADE;
NM_AREA_AVALIACAO; NR_VOLUME; NR_PAGINAS; NM_IDIOMA;
DS_RESUMO; DS_URL_TEXTO_COMPLETO; ID_PESSOA_ORIENTADOR

```

Fonte: Dados de pesquisa (2024).

A natureza dos dados armazenados nos campos expostos na Figura 6 varia entre tipos inteiro e numérico. É importante ressaltar a inserção de campos destinados à identificação tanto de Docentes quanto de Discentes. Exemplificando, o campo 'ID_PESSOA_ORIENTADOR' serve para identificar de maneira única o orientador associado a determinada produção. Este identificador permanece constante, independentemente das variações no vínculo do orientador, grau acadêmico ou Programa de Pós-Graduação. A evolução da estrutura de dados nos arquivos de 2017 a 2022 introduziu dois campos adicionais: 'IN_TCC_COM_VINCULO_PRODUCAO' e 'ID_ADD_PRODUCAO_VINCULO_CT'. O primeiro tem como propósito discernir se o trabalho possui associação com alguma produção intelectual, marcando o valor "SIM", caso positivo. Já o segundo campo, fornece os códigos de identificação das produções associadas, apresentados no formato exemplificado: '31018838 | 31018839'.

A tendência crescente na complexidade e na granularidade dos dados, especialmente no período de 2013 a 2022, sugere que houve um esforço concentrado por parte da CAPES para aprimorar a coleta e a curadoria dos dados. Esta evolução pode ser vista tanto na adição de campos mais específicos quanto na estruturação dos próprios dados, indicando uma maturidade crescente na gestão da informação. A introdução de campos que especificam vinculações entre trabalhos e outras produções intelectuais é um reflexo da necessidade contemporânea de entender o impacto e a interconexão entre diferentes pesquisas. Em um cenário acadêmico cada vez mais colaborativo e interdisciplinar, é fundamental dispor de uma

visão clara de como as pesquisas se conectam e contribuem mutuamente (Budler; Župič; Trkman, 2021).

Em resumo, o aprimoramento observado nos DataFrames DF2 não só facilita a análise e extração de insights por pesquisadores e acadêmicos, mas também reflete uma evolução no ecossistema de pesquisa do Brasil, onde a qualidade e a integridade dos dados são cada vez mais valorizadas (Sarti, 2017). Esta tendência é essencial para impulsionar a ciência baseada em evidências e garantir que as decisões tomadas, seja no âmbito acadêmico ou de políticas públicas, sejam embasadas em informações precisas e confiáveis.

4.5.1.3 Etapa 3 - Consolidação dos conjuntos de dados

Após o tratamento e a normalização, os dados são organizados em estruturas analíticas conhecidas como DataFrames, que funcionam como tabelas relacionais e estruturadas, permitindo o armazenamento, manipulação e análise eficiente das informações. Essa etapa envolve a consolidação dos diferentes arquivos de origem em conjuntos únicos, padronizados e coerentes entre si, o que facilita o cruzamento de variáveis, a aplicação de filtros e a execução de análises estatísticas e exploratórias. A etapa de consolidação dos conjuntos de dados é composta por cinco fases, a saber: Fase 1 - Integração dos *DataFrames*, Fase 2 - Agregação Preliminar dos Arquivos de Dados, Fase 3 - Acoplamento dos arquivos de 2004 a 2012 no DF1, Acoplamento dos arquivos de 2013 a 2022 no D e Fase 5 - Normalização dos dados do DF1 e DF2, descritas a seguir.

4.5.1.3.1 Fase 1 - Integração dos *DataFrames*

A integração de *DataFrames* complexos é uma etapa da análise de dados que permite uma visão mais abrangente das informações disponíveis (Armbrust *et al.*, 2015). Para a união de DF1 e DF2, foi necessário garantir que as informações fossem combinadas de maneira coesa e coerente, conforme os seguintes passos:

1. Agregação Preliminar dos Arquivos de Dados:
 - a. Antes de qualquer operação de integração, é imperativo que os arquivos individuais de DF1 e DF2 sejam amalgamados em dois conjuntos de dados únicos, respectivamente. Esse passo facilitará a subsequente operação de integração e minimizará o potencial de inconsistências (Rahm, 2000).
2. Estabelecimento do Esquema do DF3_trabalhos_BR:

- a. Posterior à agregação, as variáveis comuns entre os conjuntos de dados unificados de DF1 e DF2 devem ser identificadas e catalogadas (Rahm, 2000).
- b. Subsequentemente, as variáveis exclusivas de cada conjunto de dados serão discernidas (Rahm, 2000).
- c. O DF3_trabalhos_BR, conseqüentemente, incorporará todas essas variáveis, englobando tanto as comuns quanto as exclusivas. As variáveis que estão unicamente presentes em um conjunto de dados serão complementadas com valores nulos para as entradas pertencentes ao conjunto de dados onde a variável não é originalmente presente (Rahm, 2000).

3. Mapeamento e Correspondência de Variáveis:

- a. A correspondência entre as variáveis de DF1 e DF2, que possuem semelhanças semânticas mas nomenclaturas divergentes, deve ser meticulosamente mapeada (Rahm, 2000; Simmhan, Plale, Gannon, 2005).
- b. Variáveis singulares a um conjunto de dados serão integradas ao DF3_trabalhos_BR com entradas designadas como nulas onde aplicável. Variáveis singulares a um conjunto de dados serão integradas ao DF3_trabalhos_BR com entradas designadas como nulas onde aplicável (Rahm, 2000; Simmhan, Plale, Gannon, 2005).

4. Deliberações Acerca das Variáveis Exclusivas:

- a. A compreensão intrínseca da natureza e relevância das variáveis exclusivas em DF2 é fundamental (Rahm, 2000; Simmhan, Plale, Gannon, 2005).
- b. Entradas de DF1 serão designadas com valores nulos para tais variáveis exclusivas (Rahm, 2000; Simmhan, Plale, Gannon, 2005).

5. Validação e Revisão Subsequente à Integração:

- a. Após a integração, é mandatório conduzir uma revisão minuciosa do conjunto de dados unificado para assegurar a integridade e precisão dos dados amalgamados (Rahm, 2000).

Em síntese, o arcabouço (conjunto de elementos, regras e métodos usados no processo de integração dos dados) proposto para a integração dos conjuntos de dados DF1 e DF2 em DF3_trabalhos_BR é meticuloso, garantindo que todas as variáveis, sejam elas comuns ou exclusivas, sejam adequadamente tratadas e representadas. Tal abordagem tem o intuito de assegurar que o DF3_trabalhos_BR seja um repositório de dados robusto e integral para futuras investigações, mantendo informações históricas enquanto incorpora novas métricas e insights (que possam revelar padrões, tendências ou relações relevantes), facilitando uma compreensão mais profunda e bem-informada para futuras investigações.

4.5.1.3.2 Fase 2 - Agregação Preliminar dos Arquivos de Dados

A consolidação e integração de conjuntos de dados são etapas críticas em qualquer pesquisa que visa a análise precisa de informações distribuídas em diversos arquivos. Para a constituição do conjunto de dados, denominado DF3_trabalhos_BR, é imprescindível que haja uma avaliação criteriosa das variáveis contidas em cada *DataFrame* disponível (Cassavia *et al.*, 2014). Esta análise pode resultar na identificação de variáveis redundantes ou irrelevantes para a pesquisa, as quais poderão ser alvo de exclusão. Além disso, dada a evolução temporal e a diversidade de campos em *DataFrames* diferentes, adaptações e inserções de novas variáveis podem ser necessárias para garantir uma representatividade abrangente (Furche *et al.*, 2016).

Inicialmente, os arquivos referentes ao intervalo temporal de 2004 a 2012 são consolidados para formar o DF1. Este processo de agregação garante uma base de dados unificada para o referido período, otimizando as etapas subsequentes de manipulação e análise (Furche *et al.*, 2016). Posteriormente, com base nas observações das Figuras 5 e 6, constata-se uma heterogeneidade nas nomenclaturas dos campos entre os *DataFrames*. Para o DF1, esta discrepância nomenclatural demanda um trabalho detalhado de:

1. Renomear variáveis: A padronização dos nomes de campos é necessária para garantir a coesão e facilitar a integração posterior com outros conjuntos de dados.
2. Excluir variáveis: Campos que não agreguem valor significativo à pesquisa, seja pela redundância ou pela irrelevância do conteúdo, são criteriosamente excluídos, visando a otimização do *DataFrame*.
3. Incluir variáveis: A observação dos *DataFrames* revela que há campos presentes em conjuntos de dados subsequentes, como o DF2, que não estão

contidos no DF1. Estes campos são introduzidos no DF1, mesmo que inicialmente não contenham dados, a fim de preparar a estrutura para a integração futura e assegurar que todas as variáveis relevantes estejam representadas.

Através destes procedimentos, o DF1 e o DF2 estarão estruturados e aptos para as próximas fases da pesquisa, facilitando a integração e as análises subsequentes.

4.5.1.3.3 Fase 3 - Acoplamento dos arquivos de 2004 a 2012 no DF1

Dentro do escopo de análise científica, é imperativo assegurar a integridade e uniformidade dos conjuntos de dados, como visto anteriormente. Assim, para os registros referentes ao período de 2004 a 2012, foi realizado um procedimento de agregação, culminando na formação do *DataFrame* designado como DF1.

Antes de concretizar essa união, é preciso conduzir uma inspeção detalhada em cada arquivo para identificar possíveis células desprovidas de informação, garantindo que quaisquer lacunas sejam tratadas apropriadamente. Além disso, é de vital importância abordar a questão da codificação de caracteres. Os registros, originalmente em *Wester Europ (Windows-1252/WinLatin 1)*, necessitarão ser convertidos para o padrão UTF-8 (Sterne *et al.*, 2009; Stodder; Matters, 2016). Esta transformação assegura a consistência na representação textual, eliminando potenciais incompatibilidades ou erros de interpretação e facilitando a interoperabilidade dos dados em plataformas e ferramentas diversas.

Nesta fase do trabalho, o objetivo principal é assegurar a integridade do *DataFrame*, composto por registros anuais de 2004 a 2012. Cada um destes arquivos CSV é analisado para identificar células desprovidas de informação. Para atingir tal objetivo, propõe-se a utilização da linguagem de programação, como pode ser visto no Quadro 23, que mostra o processo para inserção do valor *NULL* para os campos vazios.

Quadro 23 – Script utilizado para inserir valor null para as células vazias nos arquivos referentes aos anos de 2004 a 2012

```
import os
import pandas as pd
# Definindo o caminho para os arquivos
path = '/content/drive/MyDrive/dados_abertos/DF1'
# Listando todos os arquivos CSV no diretório especificado
files = [f for f in os.listdir(path) if f.endswith('.csv')]
# Iterando sobre cada arquivo para tratar células vazias
for file in files:
    filepath = os.path.join(path, file)
    try:
        # Lendo o arquivo usando o delimitador ';'
        df = pd.read_csv(filepath, delimiter=';')
```

```
# Substituindo células vazias por "NULL"
df.fillna("NULL", inplace=True)
except Exception as e:
    print(f"Erro ao processar o arquivo {file}. Erro: {e}")
print("Tratamento de células vazias concluído.")
```

Fonte: Dados de pesquisa (2024).

Após a inclusão de valores nulos para os valores ausentes, a etapa subsequente é a transformação da codificação de caracteres de cada arquivo. Os dados, inicialmente codificados como *Windows-1252*, devem ser convertidos para o padrão UTF-8 (Sterne *et al.*, 2009), como indicado no Quadro 23. Esta etapa não só facilita a leitura e interpretação dos dados em diversas plataformas, mas, também, previne erros potenciais de interpretação de caracteres especiais ou acentuados. Os arquivos no diretório especificado são lidos utilizando a codificação ISO-8859-1 (ou *Windows-1252*, dependendo do conteúdo original dos arquivos). Posteriormente, cada arquivo é salvo com a codificação UTF-8, mantendo o conteúdo original, mas atualizando sua codificação (Stodder; Matters, 2016). No Quadro 24 pode-se observar o código que troca o encoding¹² de cada arquivo.

Quadro 24 – Script para troca de encoding dos arquivos referentes aos anos de 2004 a 2012

```
import os
import pandas as pd
# Definindo o caminho para os arquivos
path = '/content/drive/MyDrive/dados_abertos/DF1'
# Listando todos os arquivos CSV no diretório especificado
files = [f for f in os.listdir(path) if f.endswith('.csv')]
# Iterando sobre cada arquivo para converter a codificação
for file in files:
    filepath = os.path.join(path, file)
    try:
        # Lendo o arquivo usando a codificação 'ISO-8859-1'
        df = pd.read_csv(filepath, encoding='ISO-8859-1')
        # Salvando o arquivo de volta com a codificação 'UTF-8'
        df.to_csv(filepath, index=False, encoding='UTF-8')
    except Exception as e:
        print(f"Erro ao processar o arquivo {file}. Erro: {e}")
print("Conversão de codificação concluída.")
```

Fonte: Dados de pesquisa (2024).

No ambiente da pesquisa científica, uma vez que os dados estão em formatos padrão e codificações uniformes, o próximo desafio envolve a consolidação desses dados (ou realizar

¹² O *encoding* (ou codificação) de arquivos refere-se ao esquema que define como os caracteres são armazenados em formato digital. Diferentes tipos de encoding são usados para representar caracteres e símbolos em arquivos de texto, dependendo da língua e do sistema de escrita, além de garantir a compatibilidade com diferentes sistemas e *softwares*. Um exemplo comum de encoding é o UTF-8, que foi utilizado neste estudo por sua capacidade de representar caracteres de diferentes idiomas e sistemas de escrita.

um merge) em um conjunto único para análises integradas. Este é particularmente o caso quando lida-se com séries temporais ou dados distribuídos ao longo do tempo. A vantagem dessa abordagem é criar uma visualização holística, permitindo análises mais abrangentes. No Quadro 25 pode-se observar o script que faz o merge nos arquivos de 2004 a 2012 e transforma-os no DF1.csv.

Quadro 25 – Merge dos arquivos de 2004 a 2012

```
import os
import pandas as pd
import matplotlib.pyplot as plt
# Definindo o caminho para os arquivos
path = '/content/drive/MyDrive/CAPEs/datasets/dados_abertos/DF1'
# Listando todos os arquivos CSV no diretório especificado
files = sorted([f for f in os.listdir(path) if f.endswith('.csv')])
dfs = [] # Lista para armazenar os DataFrames
# Lendo e armazenando cada arquivo na lista
for file in files:
    year = file.split('.')[0] # Extrair o ano a partir do nome do arquivo
    df = pd.read_csv(os.path.join(path, file), encoding='UTF-8')
    df['Ano'] = year # Adicionar coluna de Ano
    dfs.append(df)
# Concatenando todos os DataFrames para criar o DF1
DF1 = pd.concat(dfs, ignore_index=True)
# Salvar DF1 como CSV
DF1.to_csv('/content/drive/MyDrive/dados_abertos/DF1/DF1.csv',
index=False, encoding='UTF-8')
# Verificações
print(f"Total de registros no DF1: {len(DF1)}")

# Contagem de registros por ano
print("\nRegistros por ano:")
print(DF1['Ano'].value_counts())
```

Fonte: Dados de pesquisa (2024).

Consequentemente à execução programática do procedimento de agregação, o conjunto de dados resultante, denominado DF1, compreende um total de 423.505 registros. Ao segmentar essa composição por temporalidade, observa-se a seguinte distribuição anual:

- 2012: 61.053 registros
- 2011: 55.554 registros
- 2010: 50.903 registros
- 2009: 50.167 registros
- 2008: 46.750 registros
- 2007: 42.819 registros
- 2006: 41.647 registros
- 2005: 39.695 registros

- 2004: 34.917 registros

Ao aferir estes valores quantitativos com os especificados na Tabela 1, confirma-se uma concordância exata entre os dados, corroborando a precisão e eficácia do procedimento de acoplamento dos conjuntos de dados individuais. Esta congruência assegura que a consolidação e subsequente construção do DF1 foi realizada com rigor metodológico, garantindo a integridade dos registros consolidados para análises subsequentes no contexto científico.

O acoplamento bem-sucedido dos arquivos em um conjunto de dados único, o DF1, não só otimiza a análise subsequente, mas também garante a integridade e consistência dos dados (Sterne *et al.*, 2009). A visualização da quantidade de campos nulos é utilizada para compreender a completude dos dados e planejar etapas de limpeza ou imputação (Stodder; Matters, 2016). Com o DF1 estabelecido, o próximo passo é aplicar um procedimento similar para os arquivos de 2013 a 2021, culminando na criação do DF2. Estas etapas preparatórias são cruciais para qualquer análise subsequente, garantindo que os dados estejam em um formato adequado e padronizado para análises mais profundas.

4.5.1.3.4 Fase 4 - Acoplamento dos arquivos de 2013 a 2022 no DF2

Com a evolução dos conjuntos de dados e a diversificação das pesquisas, os arquivos referentes aos anos de 2013 a 2022 apresentaram uma riqueza maior de detalhes em comparação com os anos anteriores (Stodder; Matters, 2016). Estes arquivos, categorizados sob a denominação DF2, representam um longo período na pós-graduação *stricto sensu* no Brasil e, conseqüentemente, exigem um tratamento meticuloso para garantir a integridade e a qualidade da informação.

Para o acoplamento deste conjunto de dados, deve-se repetir os mesmos passos que foram aplicados na Fase 4, anterior. De modo abstratizados, os passos para acoplamento de conjuntos de dados, aplicáveis a qualquer projeto, são os seguintes:

1. Preparação e organização dos conjuntos de dados: identificar e reunir os arquivos ou tabelas a serem integradas, considerando sua estrutura e relevância para a análise.
2. Inspeção e tratamento de valores ausentes: efetuar uma verificação para identificar e tratar células vazias ou valores ausentes, preenchendo-os com um valor padrão ou removendo-os conforme a necessidade.

3. Padronização de formatos e codificações: verificar e, se necessário, ajustar a codificação e o formato dos dados para um padrão comum, garantindo compatibilidade e evitando problemas de interoperabilidade.
4. Integração dos dados: usar uma técnica de acoplamento adequada (como merge ou concatenação), garantindo que os dados estejam alinhados com base em uma chave comum ou outra lógica relevante.
5. Validação do conjunto consolidado: confirmar a integridade do conjunto de dados resultante, verificando a quantidade de registros e a consistência dos dados.
6. Salvamento e preparo para análise: salvar o conjunto consolidado em um formato adequado para as análises subsequentes, certificando-se de que os dados estejam organizados e prontos para serem utilizados.

Após a aplicação desses seis passos para a operação de acoplamento, o Quadro 26 pode ser observado o script que irá inserir a informação NULL nos campos vazios do DF2.

Quadro 26 – Script utilizado para inserir valor null para as células vazias nos arquivos referentes aos anos de 2013 a 2021

```
import os
import pandas as pd
# Definindo o caminho para os arquivos
path = '/content/drive/MyDrive/dados_abertos/DF2'
# Listando todos os arquivos CSV no diretório especificado
files = [f for f in os.listdir(path) if f.endswith('.csv')]
# Iterando sobre cada arquivo para tratar células vazias
for file in files:
    filepath = os.path.join(path, file)
    try:
        # Lendo o arquivo usando o delimitador ';'
        df = pd.read_csv(filepath, delimiter=';', encoding='ISO-8859-1')
    )
        # Substituindo células vazias por "NULL"
        df.fillna('NULL', inplace=True)
    except Exception as e:
        print(f"Erro ao processar o arquivo {file}. Erro: {e}")
    print("Tratamento de células vazias concluído.")
```

Fonte: Dados de pesquisa (2024).

Os arquivos do DF2, referentes aos anos de 2013 a 2022, possivelmente foram gerados ou salvos em diferentes padrões de codificação ao longo do tempo. Estas variações podem levar a caracteres mal interpretados, especialmente quando se tratam de acentos ou caracteres especiais. Assim como feito no DF1, essa etapa reforça a importância de se trabalhar com dados limpos e consistentes, preparando o terreno para o acoplamento e análise futura dos DataFrames. Para garantir a uniformidade e a integridade dos dados dos arquivos do DF2, estes foram convertidos para a codificação 'UTF-8'. No Quadro 27 pode-se observar o *script* utilizado para alteração do encoding nos arquivos referentes aos anos de 2013 a 2022.

Quadro 27 – Script para troca de encoding dos arquivos referentes aos anos de 2013 a 2022

```
import os
import pandas as pd
# Definindo o caminho para os arquivos
path = '/content/drive/MyDrive/dados_abertos/DF1'
# Listando todos os arquivos CSV no diretório especificado
files = [f for f in os.listdir(path) if f.endswith('.csv')]
# Iterando sobre cada arquivo para converter a codificação
for file in files:
    filepath = os.path.join(path, file)
    try:
        # Lendo o arquivo usando a codificação 'ISO-8859-1'
        df = pd.read_csv(filepath, encoding='ISO-8859-1')
        # Salvando o arquivo de volta com a codificação 'UTF-8'
        df.to_csv(filepath, index=False, encoding='UTF-8')
    except Exception as e:
        print(f'Erro ao processar o arquivo {file}. Erro: {e}')
print("Conversão de codificação concluída.")
```

Fonte: Dados de pesquisa (2024).

Em uma análise aprofundada dos conjuntos de dados referentes aos anos de 2013 a 2022, foi identificada a presença de dois campos que não estão presentes em todos os arquivos: `IN_TCC_COM_VINCULO_PRODUCAO` e `ID_ADD_PRODUCAO_VINCULO_CT`. Estes campos, à primeira vista, parecem representar características específicas ou novas métricas, introduzidas a partir de 2017. No entanto, a frequência de preenchimento destes campos é relativamente baixa. Dado o contexto e a baixa representatividade destes campos, foi tomada a decisão de proceder com a remoção dos dois campos dos conjuntos de dados para os anos de 2017 a 2022. Esta escolha está alinhada com os princípios de limpeza e preparação de dados, buscando garantir a robustez e precisão das análises subsequentes (Stodder; Matters, 2016).

Uma vez tratados os pontos anteriores, a junção dos arquivos individuais dos anos de 2013 a 2022 se torna uma tarefa mais direta. Utilizando a linguagem de programação, cada arquivo é concatenado sequencialmente, respeitando-se a estrutura original dos campos e garantindo que os registros sejam mantidos de forma ordenada. Ao final deste processo, os 9 arquivos são fundidos em um único DataFrame, denominado 'DF2.csv', que representa, de forma consolidada, os registros dos anos em questão. No Quadro 28 pode ser visto o código para executar o acoplamento dos arquivos de 2013 a 2022, gerando o arquivo DF2.csv.

Quadro 28 – Merge dos arquivos de 2013 a 2022

```
import os
import pandas as pd
import matplotlib.pyplot as plt
# Definindo o caminho para os arquivos
path = '/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF2'
# Listando todos os arquivos CSV no diretório especificado
files = sorted([f for f in os.listdir(path) if f.endswith('.csv')])
dfs = [] # Lista para armazenar os DataFrames
# Lendo e armazenando cada arquivo na lista
```

```

for file in files:
    year = file.split('.')[0] # Extrair o ano a partir do nome do arquivo
    df = pd.read_csv(os.path.join(path, file), encoding='UTF-8')
    df['Ano'] = year # Adicionar coluna de Ano
    dfs.append(df)
# Concatenando todos os DataFrames para criar o DF1
DF1 = pd.concat(dfs, ignore_index=True)
# Salvar DF1 como CSV
DF1.to_csv('/content/drive/MyDrive/dados_abertos/DF1/DF1.csv',
index=False, encoding='UTF-8')
# Verificações
print(f"Total de registros no DF1: {len(DF1)}")

# Contagem de registros por ano
print("\nRegistros por ano:")
print(DF1['Ano'].value_counts())

```

Fonte: Dados de pesquisa (2024).

Com a finalidade de estender a análise e inclusão dos dados dos anos subsequentes, foi realizado um acoplamento dos conjuntos de dados dos anos 2013 a 2021, resultando no conjunto denominado DF2. Esse processo culminou na seguinte distribuição anual de registros:

- 2019: 94.503 registros
- 2018: 90.469 registros
- 2017: 85.310 registros
- 2022: 82.367 registros
- 2016: 80.278 registros
- 2020: 80.114 registros
- 2021: 79.985 registros
- 2015: 75.884 registros
- 2014: 70.668 registros
- 2013: 67.534 registros

Ao comparar esses números com os dados originais, representados na Tabela 2, confirma-se a precisão da execução dos *scripts*, refletindo a integridade e consistência dos registros incorporados. A correta consolidação em DF2 assegura que a metodologia empregada foi rigorosa e fiel à originalidade dos dados. A obtenção bem-sucedida do DF2, que contempla os registros de 2013 a 2022, complementa a base de dados anteriormente consolidada em DF1. Juntos, esses conjuntos fornecem uma visão abrangente da temporalidade do *corpus* de dados. No entanto, ao se observar a estrutura de ambos, percebe-se a existência de discrepâncias nos campos. Enquanto o DF1 possui 41 campos, o DF2 conta com até 58. Tal constatação sinaliza a necessidade de uma harmonização dos campos antes de qualquer tentativa de fundir os dois DataFrames.

Diante disso, o próximo passo é a realização de uma normalização entre os campos do DF1 e DF2, descrita a seguir, na Fase 5.

4.5.1.3.5 Fase 5 - Normalização dos dados do DF1 e DF2

À medida que há avanços na consolidação e na análise dos conjuntos de dados, surge a necessidade intrínseca de normalização, visando um acoplamento bem-sucedido e coeso (Lenzerini, 2002). As Figuras 5 e 6 ilustram a estrutura dos campos presentes do DF1 e DF2, respectivamente. Uma análise criteriosa revela algumas discrepâncias entre os dois DataFrames. No DF1, por exemplo, observamos a inclusão de múltiplos campos destinados aos nomes dos orientadores, uma característica ausente no DF2. Por outro lado, DF2 apresenta um grau maior de detalhamento em aspectos como resumos (*abstracts*), palavras-chave, participação do orientador na banca examinadora, área de expectativa de atuação, data de matrícula, denominação da categoria do orientador, entre outros.

Em virtude dessas divergências, torna-se imperativo refinar a seleção de campos que contribuirão para a composição do DataFrame final (Lenzerini, 2002). Na Tabela 3, delineiam-se os campos de DF1 que serão retidos e descartados. Para facilitar a visualização dos campos que não serão utilizados, as células correspondentes foram assinaladas com o termo “Não” e destacadas na cor vermelha.

Tabela 3 – Campos existentes no DF1

Índice	Nome do campo	Breve descrição	Será usado?
1	AnoBase	Ano Base de Coleta dos Dados.	Sim
2	CodigoPrograma	O Código do programa é gerado a partir do código da IES na Capes, já que o programa é vinculado a IES. O código do programa é único.	Sim
3	Regiao	Nome da Região da IES.	Sim
4	Uf	Sigla da Unidade da Federação da IES.	Sim
5	SiglaIes	Sigla da Instituição de Educação Superior.	Sim
6	NomeIes	Nome da Instituição de Educação Superior.	Sim
7	NomePrograma	Nome do Programa.	Sim
8	GrandeAreaCodigo	Código da Grande Área de Conhecimento a que a Produção está vinculada.	Sim
9	GrandeAreaDescricao	Nome da Grande Área de Conhecimento a que a Produção está vinculada	Sim

Índice	Nome do campo	Breve descrição	Será usado?
10	AreaConhecimentoCodigo	Código da Área de Conhecimento a que o Programa de Pós-Graduação está vinculada	Sim
11	AreaConhecimento	Nome da Área de Conhecimento a que o Programa de Pós-Graduação está vinculada	Sim
12	AreaAvaliacao	Nome da Área de Avaliação	Sim
13	DocumentoDiscente	CPF Tarjado do Discente	Sim
14	Autor	Nome do Autor da Tese ou Dissertação	Sim
15	TituloTese	Título da Tese ou Dissertação	Sim
16	Nível	Nível do Programa	Sim
17	DataDefesa	Data da Defesa da Tese ou Dissertação	Não
18	PalavrasChave	Palavras Chave da Tese ou Dissertação	Sim
19	Volume	Descrição da Biblioteca Depositária	Não
20	NumeroPaginas	Número do Volume da Produção	Não
21	BibliotecaDepositaria	Número de Páginas da Produção	Não
22	Idioma	Nome do Idioma da Produção	Sim
23	ResumoTese	Descrição do Resumo	Sim
24	LinhaPesquisa	Descrição da Linha de Pesquisa	Não
25	URLTextoCompleto	Descrição da URL do Texto Completo	Não
26	DocumentoOrientador_1	CPF Tarjado do Orientador 1	Sim
27	Orientador_1	Nome do Orientador 1	Sim
28	DocumentoOrientador_2	CPF Tarjado do Orientador 2	Não
29	Orientador_2	Nome do Orientador 2	Não
30	DocumentoOrientador_3	CPF Tarjado do Orientador 3	Não
31	Orientador_3	Nome do Orientador 3	Não
32	DocumentoOrientador_4	CPF Tarjado do Orientador 4	Não
33	Orientador_4	Nome do Orientador 4	Não
34	DocumentoCoOrientador_1	1 CPF Tarjado do Coorientador 1	Não
35	CoOrientador_1	Nome do Coorientador 1	Não
36	DocumentoCoOrientador_2	CPF Tarjado do Coorientador 2	Não
37	CoOrientador_2	Nome do Coorientador 2	Não
38	DocumentoCoOrientador_3	CPF Tarjado do Coorientador 3	Não
39	CoOrientador_3	Nome do Coorientador 3	Não

Índice	Nome do campo	Breve descrição	Será usado?
40	DocumentoCoOrientador_4	CPF Tarjado do Coorientador 4	Não
41	CoOrientador_4	Nome do Coorientador 4	Não

Fonte: Adaptado da base CAPES – Banco de Metadados.

Analogamente, a Tabela 4 faz o mesmo para o DF2, estabelecendo uma hierarquia de relevância e aplicabilidade para o escopo deste estudo, e, para também facilitar a visualização dos campos que não serão utilizados, as células foram assinaladas com o termo "Não", destacadas com a cor vermelha.

Tabela 4 – Campos existentes no DF2

Índice	Nome do campo	Breve descrição	Será usado?
1	AN_BASE	Ano Base de Coleta dos Dados	Sim
2	CD_PROGRAMA	Código de Identificação do Programa	Sim
3	NM_PROGRAMA	Nome do Programa	Sim
4	SG_ENTIDADE_ENSINO	Sigla da Entidade de Ensino	Sim
5	NM_ENTIDADE_ENSINO	Nome da Entidade de Ensino	Sim
6	ID_ADD_PRODUCAO_INTELECTUAL	Identificador do Produto no Ano Base e no Programa na Base de Dados da CAPES	Sim
7	ID_PRODUCAO_INTELECTUAL	Identificador do Produto na Base de Dados da CAPES	Sim
8	NM_PRODUCAO	Nome da Produção	Sim
9	ID_SUBTIPO_PRODUCAO	Identificador do Subtipo da Produção	Sim
10	NM_SUBTIPO_PRODUCAO	Nome do Subtipo da Produção	Sim
11	ID_AREA_CONCENTRACAO	Identificador da área de concentração	Sim
12	NM_AREA_CONCENTRACAO	Nome da área de concentração	Sim
13	ID_LINHA_PESQUISA	Identificador da Linha de Pesquisa	Não
14	NM_LINHA_PESQUISA	Nome da Linha de Pesquisa	Não
15	ID_PROJETO	Identificador do Projeto	Não
16	NM_PROJETO	Nome do Projeto	Não
17	DH_INICIO_AREA_CONC	Data e Hora do Início da Área de Concentração	Não
18	DH_FIM_AREA_CONC	Data e Hora do Fim da Área de Concentração	Não
19	DH_INICIO_LINHA	Data e Hora do Início da Linha de Pesquisa	Não

Índice	Nome do campo	Breve descrição	Será usado?
20	DH_FIM_LINHA	Data e Hora do Fim da Linha de Pesquisa	Não
21	DT_TITULACAO	Data da Titulação	Não
22	DS_PALAVRA_CHAVE	Descrição da Palavra Chave	Sim
23	DS_ABSTRACT	Descrição do Abstract	Não
24	DS_KEYWORD	Descrição do Keyword	Não
25	IN_TRABALHO_MESMA_ARE A	Indicador de Expectativa de Atuação	Não
26	NM_TP_VINCULO	Nome do Tipo de vínculo	Não
27	IN_ORIENT_PARTICIPOU_BAN CA	Indicador se o Orientador Participou da Banca	Não
28	DS_BIBLIOTECA_DEPOSITARI A	Descrição da Biblioteca Depositária	Não
29	ID_TP_EXPECTATIVA_ATUAC AO	Identificador do Tipo de Expectativa da Atuação	Não
30	NM_EXPECTATIVA_ATUACA O	Nome da Expectativa da Atuação	Não
31	ID_PESSOA_DISCENTE	Código de identificação da pessoa na base de dados da CAPES	Sim
32	NM_DISCENTE	Nome do Discente	Sim
33	DT_MATRICULA	Data da Matrícula do Discente	Sim
34	ID_GRAU_ACADEMICO	Identificador do Grau Acadêmico ao qual o discente está vinculado	Não
35	NM_GRAU_ACADEMICO	Nome do Grau Acadêmico ao qual o discente está vinculado	Não
36	NM_ORIENTADOR	Nome do Orientador	Sim
37	DS_CATEGORIA_ORIENTADO R	Descrição da Categoria do Orientador	Não
38	NM_CATEGORIA_DOCENTE	Nome da Categoria do Docente	Não
39	NM_REGIAO	Nome da Região da IES	Sim
40	SG_UF_IES	Sigla da Unidade da Federação da IES	Sim
41	NM_UF_IES	Nome da Unidade da Federação da IES	Não
42	CD_GRANDE_AREA_CONHECI MENTO	Código da Grande Área de Conhecimento a que a Produção está vinculada	Sim
43	NM_GRANDE_AREA_CONHEC IMENTO	Nome da Grande Área de Conhecimento a que a Produção está vinculada	Sim
44	CD_AREA_CONHECIMENTO	Código da Área de Conhecimento a que o	Sim

Índice	Nome do campo	Breve descrição	Será usado?
		Programa de Pós-Graduação está vinculada	
45	NM_AREA_CONHECIMENTO	Nome da Área de Conhecimento a que o Programa de Pós-Graduação está vinculada	Sim
46	CD_SUBAREA_CONHECIMENTO	Código da Subárea de Conhecimento a que o Programa de Pós-Graduação está vinculada	Não
47	NM_SUBAREA_CONHECIMENTO	Nome da Subárea de Conhecimento a que o Programa de Pós-Graduação está vinculada	Não
48	CD_ESPECIALIDADE	Código da Especialidade	Não
49	NM_ESPECIALIDADE	Nome da Especialidade	Não
50	NM_AREA_AVALIACAO	Nome da Área de Avaliação	Sim
51	NR_VOLUME	Número do Volume da Produção	Não
52	NR_PAGINAS	Número de Páginas da Produção	Não
53	NM_IDIOMA	Nome do Idioma da Produção	Sim
54	DS_RESUMO	Descrição do Resumo	Sim
55	DS_URL_TEXTO_COMPLETO	Descrição da URL do Texto Completo	Não
56	ID_PESSOA_ORIENTADOR	Código de identificação da pessoa orientador na base de dados da CAPES	Sim
57	IN_TCC_COM_VINCULO_PRODUCAO	Indicada se o TCC tem vínculo com alguma Produção Intelectual. Informação complementar não coletada no evento de 2022 para os dados de 2021.	Não
58	ID_ADD_PRODUCAO_VINCULO_CT	Representa os 'IDs' das produções intelectuais concatenadas, para caso o usuário deseje saber qual produção intelectual foi vinculada aquele TCC. Informação complementar não coletada no evento de 2022 para os dados de 2021.	Não

Fonte: Adaptado do CAPES - Banco de Metadados.

Finalizada a normalização dos dados de DF1 e DF2, passa-se à criação do DF3_trabalhos_BR.

4.5.1.4 Etapa 4 - Criação do DataFrame (DF3_trabalhos_BR)

Nesta etapa, define-se formalmente a estrutura final do DF3_trabalhos_BR, conforme apresentado na Tabela 5. Esse DataFrame consolidado serve como a base das análises subsequentes da pesquisa, integrando de modo unificado as características essenciais dos conjuntos de dados originais (DF1 e DF2). A delimitação precisa dos campos e sua

organização refletem o rigor metodológico, assegurando que as análises derivadas sejam consistentes, confiáveis e relevantes para o avanço do conhecimento, conforme alerta Lenzerini (2002).

Tabela 5 – Composição dos campos DF3_trabalhos_BR

Índice	Nome do campo	Breve descrição
1	ANO_BASE	Ano da titulação do discente.
2	CD_GRANDE_AREA_CONHECIMENTO	Código da Grande Área de Conhecimento a que a Produção está vinculada.
3	NM_GRANDE_AREA_CONHECIMENTO	Nome da Grande Área de Conhecimento a que a Produção está vinculada.
4	CD_AREA_CONHECIMENTO	Código da Área de Conhecimento a que o Programa de Pós-Graduação está vinculado.
5	NM_AREA_CONHECIMENTO	Nome da Área de Conhecimento a que o Programa de Pós-Graduação está vinculado.
6	NM_AREA_AVALIACAO	Nome da área de Avaliação a que o Programa de Pós-Graduação está vinculado quando do envio e homologação dos dados.
7	CD_PROGRAMA	O Código do programa é gerado a partir do código da IES na Capes, já que o programa é vinculado a IES. O código do programa é único.
8	NM_PROGRAMA	Nome do Programa ao qual a produção está vinculada. Nome do Programa de Pós-Graduação, informado pela Instituição e aprovado pelo Ministério da Educação.
9	SG_ENTIDADE_ENSINO	Sigla da IES (Instituição de Ensino Superior) a qual a Tese ou Dissertação está vinculada.
10	NM_ENTIDADE_ENSINO	Nome da IES (Instituição de Ensino Superior) a qual a Tese ou Dissertação está vinculada.
11	NM_REGIAO	Nome da Região onde se encontra a IES.
12	SG_UF_IES	Sigla da UF onde se encontra a IES.
13	ID_ADD_PRODUCAO_INTELCTUAL	Identificação da Tese ou Dissertação no Ano Base e no Programa na Base de Dados da CAPES.
14	ID_PRODUCAO_INTELCTUAL	Identificador da Tese ou Dissertação na Base de Dados da CAPES.
15	NM_PRODUCAO	Título da Tese ou Dissertação
16	DS_PALAVRA_CHAVE	Palavras Chave da Tese ou Dissertação.
17	DS_RESUMO	Resumo da Produção.
18	ID_GRAU_ACADEMICO	Código de identificação do grau acadêmico ao qual o discente está vinculado em 31 de dezembro do ano base.

Índice	Nome do campo	Breve descrição
19	NM_GRAU_ACADEMICO	Nome do grau acadêmico ao qual o discente está vinculado em 31 de dezembro do ano base.
20	NM_IDIOMA	Idioma da Produção.
21	ID_PESSOA_ORIENTADOR	Identifica do orientador vinculado a produção enquanto pessoa na base de dados da Capes em relação ao seu cadastro de Pessoa Física na Receita Federal. Tal identificador é único, independente do vínculo da pessoa enquanto orientador, grau acadêmico e Programa de Pós-Graduação.
22	NM_ORIENTADOR	Nome do Orientador.
23	ID_PESSOA_DISCENTE	Identifica o discente enquanto pessoa na base de dados da Capes em relação ao seu cadastro de Pessoa Física na Receita Federal. Tal identificador é único, independente do vínculo da pessoa enquanto discente, grau acadêmico e Programa de Pós-Graduação.
24	NM_DISCENTE	Nome do discente. Em geral, corresponde ao nome da pessoa na base de dados da Receita Federal, salvo atualizações nela que não tenham refletido na base de dados da Capes.

Fonte: Adaptado do CAPES - Banco de Metadados.

A Tabela 5 apresenta os campos definidos para o DF3_trabalhos_BR, destacando na cor azul aqueles que são exclusivos do DF2, que exigiram a inserção de códigos específicos durante a migração. Com o formato e a estrutura do DF3_trabalhos_BR estabelecidos, a próxima ação envolve a criação do arquivo DF3.csv, conforme ilustrado no Quadro 29, seguida da migração sequencial dos registros de DF1 e DF2 para esse novo conjunto.

Quadro 29 – Criação do arquivo DF3.csv

```
import os
import pandas as pd
# Definindo o caminho para salvar o arquivo DF3.csv
path = '/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF3.csv'
# Lista dos campos desejados
fields = [
'ANO_BASE','CD_GRANDE_AREA_CONHECIMENTO','NM_GRANDE_AREA_CONHECIMENTO',
'CD_AREA_CONHECIMENTO','NM_AREA_CONHECIMENTO','NM_AREA_AVALIACAO',
'CD_PROGRAMA','NM_PROGRAMA','SG_ENTIDADE_ENSINO','NM_ENTIDADE_ENSINO',
'NM_REGIAO','SG_UF_IES','ID_ADD_PRODUCAO_INTELECTUAL','ID_PRODUCAO_INTELECTUAL','NM_PRODUCAO','DS_P
ALAVRA_CHAVE','DS_RESUMO','ID_GRAU_ACADEMICO','NM_GRAU_ACADEMICO','NM_IDIOMA','ID_PESSOA_ORIENTA
DOR','NM_ORIENTADOR','ID_PESSOA_DISCENTE','NM_DISCENTE'
]
# Criando um DataFrame vazio com os campos desejados
df3 = pd.DataFrame(columns=fields)
# Salvando o DataFrame vazio como DF3.csv
df3.to_csv(path, sep=';', encoding='UTF-8', index=False)
# Verificando se o arquivo foi criado corretamente
df3_created = pd.read_csv(path, sep=';', encoding='UTF-8')
df3_created.head()
```

Fonte: Dados de pesquisa (2024)

A integridade e a consistência dos dados são aspectos fundamentais desse processo. Por isso, após a migração, faz-se necessária a execução de verificações detalhadas a fim de assegurar que a totalidade dos registros foi corretamente migrada e que a estrutura final do DF3_trabalhos_BR atende aos parâmetros metodológicos definidos. No que se refere à migração do DF1 para o DF3_trabalhos_BR, destaca-se a necessidade de gerar identificadores únicos para os campos ID_ADD_PRODUCAO_INTELECTUAL e ID_PRODUCAO_INTELECTUAL. Considerando a estrutura atual dos dados, esses identificadores serão gerados com valores superiores a 900000000, evitando sobreposições ou conflitos com os registros provenientes do DF2. Além disso, é realizada uma transformação específica no campo ID_GRAU_ACADEMICO, convertendo os valores descritivos do campo NM_GRAU_ACADEMICO em códigos numéricos no DF3_trabalhos_BR. A estratégia adotada será: Mestrado → Código 27; Doutorado → Código 28; Profissionalizante → Código 31. Essa codificação não apenas melhora o armazenamento e a consulta dos dados, mas também padroniza a representação dos graus acadêmicos, facilitando análises subsequentes.

Em síntese, essas transformações e migrações, embora tecnicamente desafiadoras, são essenciais para a construção de um conjunto de dados estruturado, coeso e preparado para análises aprofundadas, atendendo ao rigor e à precisão exigidos em pesquisas acadêmicas de alto nível. A seguir apresenta-se a Fase 1, da Consolidação dos DataFrames DF1 e DF2 em DF3_trabalhos_BR, composta por dois passos, a saber: de acoplamento do DF1 e do DF2 ao DF3_trabalhos_BR, respectivamente.

4.5.1.4.1 Fase 1 - Consolidação dos DataFrames DF1 e DF2 em DF3_trabalhos_BR

Nesta pesquisa, a proposta é a consolidação dos conjuntos de dados DF1 e DF2, cada um com suas especificidades, em um único conjunto denominado DF3_trabalhos_BR. Esse processo não se limita à simples agregação de dados, mas busca também sua harmonização, assegurando um conjunto coeso, uniforme e livre de redundâncias, conforme apontado na literatura (Lenzerini, 2002; Rahm, 2000; Simmhan; Plale; Gannon, 2005).

A escolha de iniciar essa fusão pelo DF1 justifica-se por sua estrutura e pelo volume de dados que apresenta. Após a transferência dos registros do DF1 para o DF3_trabalhos_BR, é realizada uma verificação minuciosa para garantir que a integração foi concluída de maneira íntegra e sem perda de registros (Wickham; Grolemond, 2017). Em seguida, esse mesmo procedimento é replicado para o DF2.

A inserção sequencial — primeiro DF1, depois DF2 — permite um controle refinado e facilita a identificação e correção de eventuais inconsistências durante o processo. Após a conclusão da transferência de ambos os conjuntos de dados, é conduzida uma verificação global no DF3_trabalhos_BR, para assegurar que todos os registros originais estejam devidamente representados. Essa etapa final garante a integridade e a completude do DF3_trabalhos_BR, preparando-o para as análises subsequentes de forma estruturada e confiável.

Para operacionalizar esse processo, a Fase 1 é dividida nos seguintes passos: 1) Passo 1: Acoplamento do DF1 ao DF3_trabalhos_BR – Transferência dos registros do DF1 para o DF3, seguida da verificação de integridade; 2) Passo 2: Acoplamento do DF2 ao DF3_trabalhos_BR – Integração dos registros do DF2 ao DF3_trabalhos_BR, com checagem final para garantir a coerência e a completude do conjunto consolidado.

Passo 1: Acoplamento DF1 ao DF3_trabalhos_BR

O processo de consolidação dos conjuntos de dados é conduzido por meio de um script em *Python*, desenvolvido especificamente para essa finalidade. Esse script, ilustrado no Quadro 30, é responsável por ler o arquivo DF1.csv, aplicar as transformações necessárias e, por fim, salvar os dados processados no arquivo DF3.csv.

Quadro 30 – Script para acoplamento do DF1 ao DF3_trabalhos_BR

```
import pandas as pd
# Caminho para os arquivos
path_df1 = '/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF1.csv'
path_df3 = '/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF3.csv'
# Ler a estrutura do arquivo DF3 para entender a ordem das colunas
df3_structure = pd.read_csv(path_df3, delimiter=';', encoding='UTF-8', nrows=0)
df3_columns = df3_structure.columns.tolist()
# Ler o DF1
df1 = pd.read_csv(path_df1, delimiter=';', encoding='UTF-8')
# Renomear as colunas conforme especificado
column_mapping = {"AnoBase": "ANO_BASE", "GrandeAreaCodigo":
"CD_GRANDE_AREA_CONHECIMENTO", "GrandeAreaDescricao":
"NM_GRANDE_AREA_CONHECIMENTO", "AreaConhecimentoCodigo": "CD_AREA_CONHECIMENTO", "AreaConhecimento":
"NM_AREA_CONHECIMENTO",
"AreaAvaliacao": "NM_AREA_AVALIACAO", "CodigoPrograma": "CD_PROGRAMA",
"NomePrograma": "NM_PROGRAMA", "SiglaIes": "SG_ENTIDADE_ENSINO",
"NomeIes": "NM_ENTIDADE_ENSINO", "Regiao": "NM_REGIAO",
"UF": "SG_UF_IES", "TituloTese": "NM_PRODUCAO", "PalavrasChave": "DS_PALAVRA_CHAVE", "ResumoTese":
"DS_RESUMO", "Nivel": "NM_GRAU_ACADEMICO", "Idioma": "NM_IDIOMA", "DocumentoOrientador_1":
"ID_PESSOA_ORIENTADOR", "Orientador_1": "NM_ORIENTADOR", "DocumentoDiscente": "ID_PESSOA_DISCENTE",
"Autor": "NM_DISCENTE"
}
df1 = df1.rename(columns=column_mapping)
# Adicionar os valores incrementados nos campos ID_ADD_PRODUCAO_INTELCTUAL e ID_PRODUCAO_INTELCTUAL
id_start = 900000000
df1["ID_ADD_PRODUCAO_INTELCTUAL"] = range(id_start, id_start + len(df1))
df1["ID_PRODUCAO_INTELCTUAL"] = range(id_start + len(df1), id_start + 2*len(df1))
# Mapear os valores de NM_GRAU_ACADEMICO para ID_GRAU_ACADEMICO
grau_mapping = {
"Mestrado": 27,
"Doutorado": 28,
"Profissionalizante": 31,
```

```

"Mestrado e Profissionalizante": 31
}
df1["ID_GRAU_ACADEMICO"] = df1["NM_GRAU_ACADEMICO"].map(grau_mapping)
# Reordenar e filtrar as colunas do DF1 de acordo com a estrutura do DF3
df1 = df1[df3_columns]
# Salvar o DF1 transformado como DF3
df1.to_csv(path_df3, sep=';', encoding='UTF-8', index=False)

```

Fonte: Dados de pesquisa (2024)

A abordagem programática assegura precisão na transferência e transformação dos registros, preservando a integridade e consistência dos dados durante a fusão (Wickham; Grolemund, 2017). Além disso, a automação reduz erros manuais e garante a replicabilidade do processo, aplicando modificações de forma sistemática e padronizada.

Após a consolidação dos conjuntos de dados no DF3_trabalhos_BR, torna-se necessária uma análise preliminar para verificar sua integridade e estrutura. Essa etapa não apenas confirma o sucesso da integração, mas também identifica eventuais inconsistências ou lacunas que possam impactar análises futuras. Utilizando a biblioteca Pandas, são aplicadas funções que permitem uma compreensão sobre a qualidade do conjunto de dados (Wickham; Grolemund, 2017). A função .shape, conforme descrito por McKinney (2012), é utilizada para retornar a dimensionalidade do DataFrame, revelando um total de 423.505 linhas e 24 colunas.

A Tabela 06 apresenta o retorno da função .info(), que, segundo McKinney (2012), fornece um resumo conciso do DataFrame, ajudando a identificar colunas com valores ausentes e a compreender os tipos de dados.

Tabela 6 – Resultado do comando info() da biblioteca Pandas

Índice	Coluna	Contador	Tipo
0	ANO_BASE	423505	int64
1	CD_GRANDE_AREA_CONHECIMENTO	423505	int64
2	NM_GRANDE_AREA_CONHECIMENTO	423505	object
3	CD_AREA_CONHECIMENTO	423505	int64
4	NM_AREA_CONHECIMENTO	423505	object
5	NM_AREA_AVALIACAO	423505	object
6	CD_PROGRAMA	423505	object
7	NM_PROGRAMA	423505	object
8	SG_ENTIDADE_ENSINO	423505	object
9	NM_ENTIDADE_ENSINO	423505	object
10	NM_REGIAO	423505	object
11	SG_UF_IES	423505	object

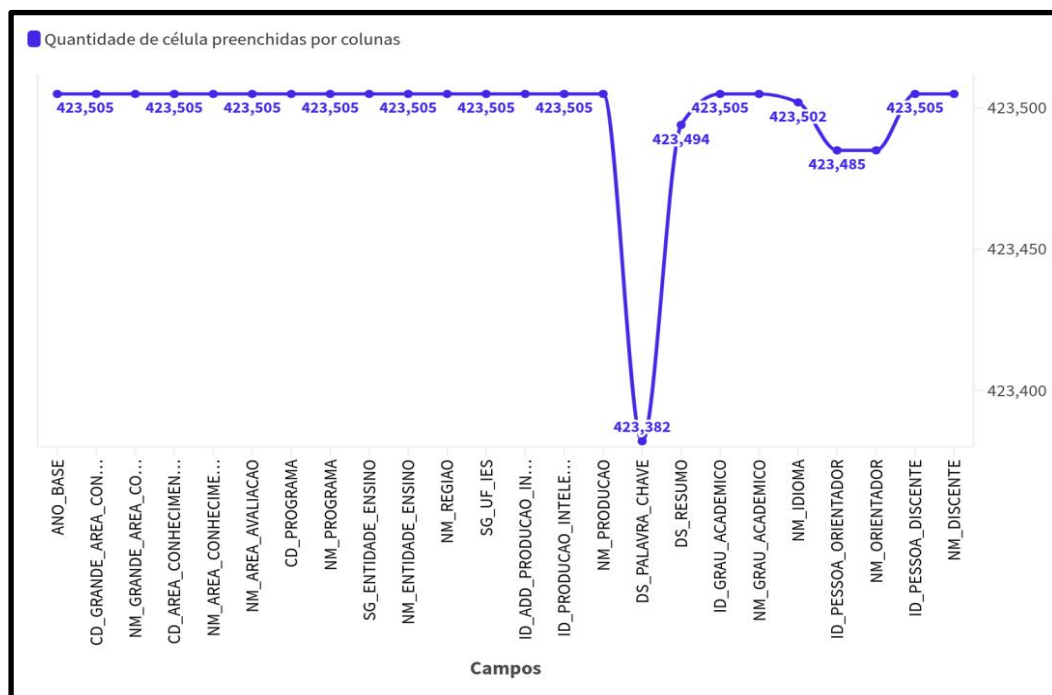
Índice	Coluna	Contador	Tipo
12	ID_ADD_PRODUCAO_INTELECTUAL	423505	int64
13	ID_PRODUCAO_INTELECTUAL	423505	int64
14	NM_PRODUCAO	423505	object
15	DS_PALAVRA_CHAVE	423382	object
16	DS_RESUMO	423494	object
17	ID_GRAU_ACADEMICO	423505	int64
18	NM_GRAU_ACADEMICO	423505	object
19	NM_IDIOMA	423502	object
20	ID_PESSOA_ORIENTADOR	423485	object
21	NM_ORIENTADOR	423485	object
22	ID_PESSOA_DISCENTE	423505	object
23	NM_DISCENTE	423505	object

Fonte: Dados de pesquisa (2024)

A análise mostrou que o campo Contador atinge o valor máximo de 423.505 registros, enquanto o campo DS_RESUMO apresenta 423.494 registros, indicando 11 células vazias. Essa verificação detalhada permite a identificação precoce de possíveis inconsistências, assegurando a qualidade do conjunto de dados consolidado.

A Gráfico 6 ilustra a quantidade de células preenchidas por coluna, evidenciando que a maioria dos campos contém 423.505 registros, enquanto alguns apresentam números ligeiramente inferiores.

Gráfico 6 – Quantidade de células preenchidas por colunas no DF3_trabalhos_BR após o acoplamento do DF1



Fonte: Dados de pesquisa (2024)

Ao analisar a Gráfico 6, observa-se que as colunas DS_PALAVRA_CHAVE, DS_RESUMO, NM_IDIOMA, ID_PESSOA_ORIENTADOR e NM_ORIENTADOR apresentam valores ausentes. A coluna DS_PALAVRA_CHAVE, possui 423.382 registros, e a coluna DS_RESUMO, possui 423.494 registros, indicando um total de 123 e 11 valores ausentes, respectivamente. Já os campos NM_IDIOMA, ID_PESSOA_ORIENTADOR e NM_ORIENTADOR apresentam pequenas variações em relação ao total esperado, o que requer uma avaliação para mitigar impactos nas análises posteriores.

Essas lacunas podem comprometer a precisão e confiabilidade das análises subsequentes, tornando fundamental compreender sua origem e definir estratégias adequadas para tratá-las. Métodos de imputação ou outras abordagens devem ser considerados antes de avançar para análises mais aprofundadas, conforme alertado por Sterne *et al.* (2009).

Com o bem-sucedido acoplamento dos dados do DF1 ao DF3_trabalhos_BR, o próximo passo é a integração do DF2 ao DF3_trabalhos_BR.

Passo 2: Acoplamento DF2 ao DF3_trabalhos_BR

A integração do DF2 para DF3_trabalhos_BR, embora possa ser intrinsicamente mais demorada devido ao tamanho substancialmente maior do DF2 — que é aproximadamente três vezes mais pesado em GB em relação ao DF1 —, é menos complexa em termos de transformações. Isso ocorre porque os campos entre DF2 e DF3_trabalhos_BR estão

alinhados, o que elimina a necessidade de modificações adicionais ou cálculos durante a integração, conforme mostra o Quadro 31.

Quadro 31 – Script para acoplamento do DF2 ao DF3_trabalhos_BR

```
import pandas as pd
# Caminhos para os arquivos
path_df2 = '/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF2.csv'
path_df3 = '/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF3.csv'
# Colunas desejadas para o DF2 e DF3
desired_columns_df2 = [
    "AN_BASE", "CD_GRANDE_AREA_CONHECIMENTO", "NM_GRANDE_AREA_CONHECIMENTO",
    "CD_AREA_CONHECIMENTO", "NM_AREA_CONHECIMENTO", "NM_AREA_AVALIACAO",
    "CD_PROGRAMA", "NM_PROGRAMA", "SG_ENTIDADE_ENSINO", "NM_ENTIDADE_ENSINO",
    "NM_REGIAO", "SG_UF_IES", "ID_ADD_PRODUCAO_INTELCTUAL",
    "ID_PRODUCAO_INTELCTUAL", "NM_PRODUCAO", "DS_PALAVRA_CHAVE",
    "DS_RESUMO", "ID_GRAU_ACADEMICO", "NM_GRAU_ACADEMICO", "NM_IDIOMA",
    "ID_PESSOA_ORIENTADOR", "NM_ORIENTADOR", "ID_PESSOA_DISCENTE", "NM_DISCENTE"
]
# Ler o DF2 e selecionar apenas as colunas desejadas
df2 = pd.read_csv(path_df2, delimiter=';', encoding='UTF-8', usecols=desired_columns_df2)
# Renomear a coluna "AN_BASE" de DF2 para "ANO_BASE"
df2 = df2.rename(columns={"AN_BASE": "ANO_BASE"})
# Tratar os campos vazios ou com apenas espaços, substituindo-os por "null"
df2 = df2.replace(["", " "], 'null')
# Ler o DF3 existente
df3 = pd.read_csv(path_df3, delimiter=';', encoding='UTF-8')
# Acoplar os dados do DF2 ao final do DF3
df3 = pd.concat([df3, df2], ignore_index=True)
# Salvar o novo DF3 consolidado
df3.to_csv(path_df3, sep=';', encoding='UTF-8', index=False)
```

Fonte: Dados de pesquisa (2024).

A harmonia entre nos atributos de DF2 e DF3_trabalhos_BR torna o processo de acoplamento mais direto, garantindo uma consolidação mais fluida e acurada dos conjuntos de dados.

Com o processo de acoplamento concluído, o *DataFrame* final consolidado apresenta uma dimensionalidade de (1232870, 24), algumas das funções da biblioteca Pandas foram executadas, como recomendações de McKinney (2012). Esta robusta estrutura, que ocupa aproximadamente 2,8GB em disco, reflete o árduo trabalho de coleta e organização de informações. Vale ressaltar que, do universo original do DF2, que abrigava 58 campos distintos, somente 24 deles foram transportados e incorporados ao DF3_trabalhos_BR. Esta decisão estratégica de seleção garantiu uma maior eficiência na análise subsequente.

Com a conclusão desse processo, o *DataFrame* final (DF3_trabalhos_BR) consolidado apresenta uma dimensionalidade de (1.148.250, 24). Para validar sua integridade estrutural, são executadas funções da biblioteca Pandas, conforme as recomendações de McKinney (2012). Essa estrutura, que ocupa aproximadamente 2,8GB em disco, reflete o árduo trabalho de coleta e organização de informações.

É importante destacar que, dos 58 campos distintos originalmente presentes no DF2, apenas 24 foram incorporados ao DF3_trabalhos_BR. Essa seleção se justifica porque evita a inclusão de variáveis com alto índice de valores ausentes ou baixa variabilidade, garantindo que apenas informações úteis sejam consideradas. Assim, a exclusão de campos como ID_LINHA_PESQUISA, NM_LINHA_PESQUISA, ID_PROJETO e NM_PROJETO, bem como as datas de início e fim das áreas de concentração e linhas de pesquisa, foi motivada pelo fato de que esses campos de dados somente começaram a existir após 2013, e estavam praticamente todos vazios, e não havia fontes complementares para preencher as lacunas existentes. Isso reforçou a necessidade de excluir variáveis com alto índice de valores ausentes e baixa disponibilidade temporal, garantindo que apenas informações completas e confiáveis sejam incorporadas ao DF3_trabalhos_BR, sem comprometer a qualidade das análises subsequentes. Essa seleção estratégica tornou possível melhorar as análises subsequentes, garantindo um conjunto de dados mais enxuto e direcionado aos objetivos da pesquisa.

Tabela 7 – Resultado do comando info() da biblioteca Pandas

Coluna	Contador	Tipo
ANO_BASE	1232870	int64
CD_GRANDE_AREA_CONHECIMENTO	1232854	int64
NM_GRANDE_AREA_CONHECIMENTO	1232849	object
CD_AREA_CONHECIMENTO	1232850	int64
NM_AREA_CONHECIMENTO	1232846	object
NM_AREA_AVALIACAO	1232839	object
CD_PROGRAMA	1232870	object
NM_PROGRAMA	1232870	object
SG_ENTIDADE_ENSINO	1232870	object
NM_ENTIDADE_ENSINO	1232870	object
NM_REGIAO	1232859	object
SG_UF_IES	1232859	object
ID_ADD_PRODUCAO_INTELECTUAL	1232870	int64
ID_PRODUCAO_INTELECTUAL	1232870	int64
NM_PRODUCAO	1232870	object
DS_PALAVRA_CHAVE	1232692	object
DS_RESUMO	1232743	object

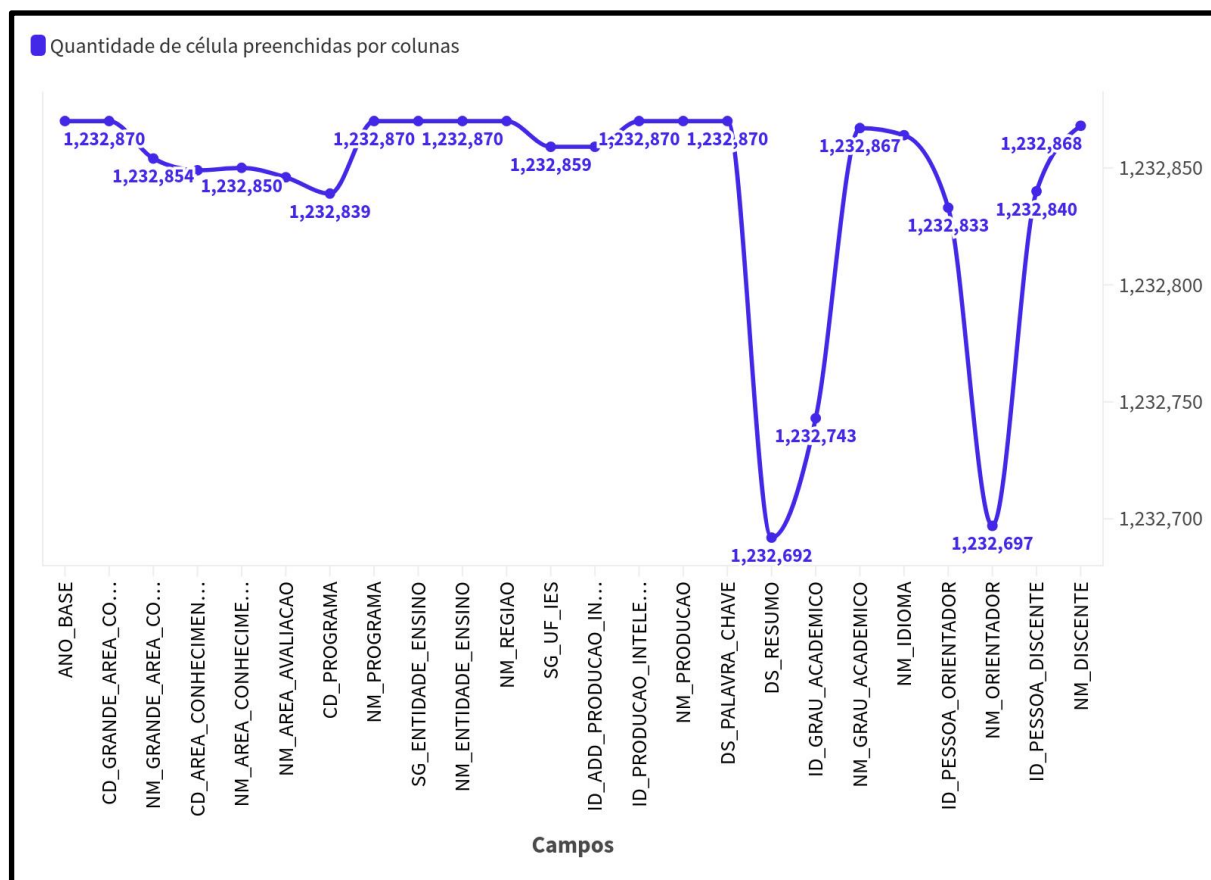
Coluna	Contador	Tipo
ID_GRAU_ACADEMICO	1232867	int64
NM_GRAU_ACADEMICO	1232864	object
NM_IDIOMA	1232833	object
ID_PESSOA_ORIENTADOR	1232697	object
NM_ORIENTADOR	1232840	object
ID_PESSOA_DISCENTE	1232868	object
NM_DISCENTE	1232866	object

Fonte: Dados de pesquisa (2024).

Ao examinar a Tabela 7, observa-se que alguns campos apresentam contagens inferiores a 1.232.870 registros, indicando a presença de valores ausentes. Essa característica, conforme Sterne *et al.* (2009), reforça a necessidade de tratamento adequado dessas lacunas para evitar impactos nas análises futuras.

Uma representação visual desse fenômeno pode ser observada no Gráfico 6, permitindo uma compreensão mais intuitiva das discrepâncias no conjunto de dados. Como poderá ser visto, em comparação com a Gráfico 7, nota-se uma variação maior na linha traçada, evidenciando a necessidade de atenção especial ao tratamento desses valores faltantes.

Gráfico 7 – Quantidade de células preenchidas por colunas no DF3_trabalhos_BR após o acoplamento do DF2



Fonte: Dados de pesquisa (2024).

Com o conjunto de dados consolidado e adequadamente preparado, as próximas ações envolvem uma análise exploratória do DF3_trabalhos_BR, aprofundando a análise a cada campo, buscando revelar padrões e tendências, para uma melhor compreensão dos dados. Além disso, a visualização de dados permitirá uma visão holística dos dados, facilitando a comunicação e compreensão dos resultados obtidos (Heer; Shneiderman, 2012).

4.5.1.5 Etapa 5 - Análise exploratória de dados (DF3_trabalhos_BR)

A análise exploratória de dados (do inglês, Exploratory data analysis (EDA)), é uma abordagem sistemática para examinar conjuntos de dados, geralmente para identificar padrões, anomalias e relações entre as variáveis. A EDA é uma etapa fundamental em qualquer pesquisa de dados, pois fornece informações para ações posteriores de modelagem e análise (Tukey, 1977). Dada a riqueza e complexidade dos Dados Abertos da CAPES, torna-se imprescindível conduzir uma análise exploratória abrangente nos dados já processados na presente pesquisa.

O processo de análise exploratória começa com um entendimento inicial do conjunto de dados, seguido por verificações de integridade e completude, referências a trabalhos anteriores, e então mergulha profundamente nas nuances dos dados. Os procedimentos são divididos em cinco fases, a saber: Fase 1: Entendimento Inicial dos Dados, Fase 2: Verificação de Valores Ausentes, Fase 3: Entendimento profundo dos dados, Fase 4: Preparação dos dados e Fase 5: Análise exploratória profunda dos dados. A seguir, apresenta-se o desenvolvimento de cada uma dessas cinco fases.

4.5.1.5.1 Fase 1 - Entendimento Inicial dos Dados

Antes de qualquer análise, é preciso entender a natureza e a estrutura dos dados com os quais se está trabalhando. Isso envolve verificar o tamanho do conjunto de dados, os tipos de variáveis presentes e uma inspeção inicial dos primeiros e últimos registros, conforme estipulado por Peng, Lee e Ingersoll (2002). Nesta ação é utilizada a linguagem de programação em *Python* e bibliotecas da mesma, para obter essas informações iniciais, garantindo uma compreensão clara do escopo e da estrutura dos dados (Lopes *et al.*, 2019).

Segundo Peng, Lee e Ingersoll (2002), qualquer questão de pesquisa exige uma análise criteriosa. Para aprofundar o entendimento dos dados, será executado um *script* em *Python*, ilustrado no Quadro 32, utilizando bibliotecas específicas cujas funções já foram detalhadas anteriormente.

Quadro 32 – Script utilizado para entendimento inicial dos dados do DF3_trabalhos_BR

```
import pandas as pd
# 1. Carregar o arquivo CSV
file_path = '/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF3.csv'
df = pd.read_csv(file_path, sep=";")
# 2. Obter as dimensões do conjunto de dados
print("Dimensões do conjunto de dados:", df.shape)
# 3. Exibir informações sobre as colunas
print("\nInformações sobre as colunas:")
print(df.info())
# 4. Exibir estatísticas descritivas das colunas numéricas
print("\nEstatísticas descritivas:")
print(df.describe())
# 5. Contar o número de valores únicos em cada coluna
print("\nNúmero de valores únicos por coluna:")
print(df.nunique())
```

Fonte: Dados de pesquisa (2024).

Conforme discutido anteriormente, o DF3_trabalhos_BR contém 1.232.870 registros distribuídos em 24 colunas. A Tabela 7 fornece uma visão detalhada da estrutura do conjunto de dados, destacando a quantidade de registros preenchidos em cada coluna. Das 24 colunas: 3 são do tipo int64, representando valores inteiros; 3 são do tipo float64, geralmente usados

para códigos ou índices decimais; 18 são do tipo *object*, categoria utilizada pelo Pandas para armazenar *strings* ou uma mistura de valores numéricos e não numéricos.

Para facilitar a manipulação dos dados, é desenvolvido um script adicional que adiciona ao DF3_trabalhos_BR uma coluna chamada "índice", atribuindo um identificador único a cada registro. Esse recurso permite localizar e recuperar registros específicos com mais facilidade, sendo particularmente útil para ajustes e análises pontuais (Lin *et al.*, 2021). O código correspondente está ilustrado no Quadro 33.

Quadro 33 – Script para inserir coluna de índice único no DF3_trabalhos_BR

```
import pandas as pd
# Carregar o arquivo CSV
file_path = '/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF3.csv'
df = pd.read_csv(file_path, sep=";")
# Criando a coluna 'índice' com números únicos
df['índice'] = range(1, len(df) + 1)
# Reordenando as colunas para que 'índice' seja a primeira coluna
colunas = ['índice'] + [col for col in df if col != 'índice']
df = df[colunas]
# Salvando o DataFrame modificado
caminho_salvo = '/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF3_com_indice.csv'
df.to_csv(caminho_salvo, sep=';', index=False)
```

Fonte: Dados de pesquisa (2024).

Embora o *DataFrame* tenha 1.232.870 linhas, algumas colunas apresentam valores ausentes. A coluna ANO_BASE contém dados para todas as linhas, enquanto NM_GRANDE_AREA_CONHECIMENTO possui 1.232.849 registros preenchidos, indicando a ausência de alguns valores. A existência de dados ausentes pode impactar os resultados e interpretações futuras. Por isso, em etapas posteriores, será necessário definir a melhor abordagem para tratá-los, seja por ***imputação*, exclusão ou outras técnicas (Nongthombam, Sharma, 2021).

4.5.1.5.2 Fase 2 - Verificação de Valores Ausentes

Dados ausentes ou nulos podem influenciar os resultados da análise. Portanto, identificar e tratar esses valores é fundamental (Nongthombam, Sharma, 2021). Os campos nulos ou vazios devem ser removidos ou ajustados, de acordo com conjunto de dados, como destacado por Nongthombam e Sharma (2021), e também devem ser consideradas estratégias adequadas para lidar com esses campos, seja imputando valores, excluindo registros ou usando técnicas mais avançadas.

Valores ausentes em conjuntos de dados são comuns e podem ocorrer por falhas na coleta, erros na entrada ou ausência intencional de informações (Nongthombam, Sharma,

2021). Durante a pesquisa, é preciso tratar esses valores adequadamente, pois podem influenciar os resultados. A ausência de dados não é necessariamente um problema, mas entender sua origem e impacto é fundamental (Little *et al.*, 2012).

Em alguns casos, a falta de um valor pode ser informativa, como em pesquisas onde a ausência de resposta pode indicar recusa. No entanto, conforme Little *et al.* (2012), recomenda-se a realização de análises de sensibilidade para verificar a robustez das conclusões frente a distintas suposições acerca dos dados ausentes. Em outros contextos, a ausência pode ser apenas um erro ou omissão.

A escolha da estratégia para lidar com esses valores depende da natureza dos dados e dos objetivos da análise. Métodos comuns incluem exclusão, imputação, substituição, interpolação ou manutenção do campo vazio (Emmanuel *et al.*, 2021; Little *et al.*, 2012). O resultado do comando `df.isnull().sum()` é apresentado na Tabela 8.

Tabela 8 – Resultado do comando `isnull().sum()` da biblioteca Pandas

Índice	Campo	Valores nulos
1	ANO_BASE	0
2	CD_GRANDE_AREA_CONHECIMENTO	16
3	NM_GRANDE_AREA_CONHECIMENTO	21
4	CD_AREA_CONHECIMENTO	20
5	NM_AREA_CONHECIMENTO	24
6	NM_AREA_AVALIACAO	31
7	CD_PROGRAMA	0
8	NM_PROGRAMA	0
9	SG_ENTIDADE_ENSINO	0
10	NM_ENTIDADE_ENSINO	0
11	NM_REGIAO	11
12	SG_UF_IES	11
13	ID_ADD_PRODUCAO_INTELECTUAL	0
14	ID_PRODUCAO_INTELECTUAL	0
15	NM_PRODUCAO	0
16	DS_PALAVRA_CHAVE	178
17	DS_RESUMO	127
18	ID_GRAU_ACADEMICO	3
19	NM_GRAU_ACADEMICO	6

Índice	Campo	Valores nulos
20	NM_IDIOMA	37
21	ID_PESSOA_ORIENTADOR	173
22	NM_ORIENTADOR	30
23	ID_PESSOA_DISCENTE	2
24	NM_DISCENTE	4

Fonte: Dados de pesquisa (2024).

A Tabela 8 mostra que a análise de dados ausentes garante a integridade dos resultados. Utilizando a biblioteca Pandas, foram identificadas colunas no DF3_trabalhos_BR com registros não preenchidos, conforme mostrado na Tabela 08. A coluna CD_GRANDE_AREA_CONHECIMENTO possui 16 valores ausentes, enquanto sua correspondente, NM_GRANDE_AREA_CONHECIMENTO, apresenta 21, sugerindo possíveis inconsistências na correspondência entre códigos e nomes. De forma similar, CD_AREA_CONHECIMENTO e NM_AREA_CONHECIMENTO têm 20 e 24 valores ausentes, respectivamente. As colunas NM_REGIAO e SG_UF_IES apresentam 11 registros vazios cada, indicando que a correção de uma pode impactar a outra. O mesmo ocorre com ID_GRAU_ACADEMICO e NM_GRAU_ACADEMICO, com 3 e 6 valores ausentes. Além disso, campos como ID_PESSOA_ORIENTADOR, NM_ORIENTADOR, ID_PESSOA_DISCENTE e NM_DISCENTE também contêm dados faltantes, reforçando a necessidade de investigação.

O tratamento desses registros deve considerar as relações entre as colunas e a origem dos dados ausentes para definir a abordagem mais adequada, seja imputação, exclusão ou outra técnica (Little *et al.*, 2012).

4.5.1.5.3 Fase 3 - Entendimento profundo dos dados

Nesta fase é realizada a compreensão dos dados, investigando estatísticas, distribuições e possíveis relações entre variáveis, sendo necessário referir-se às colunas, indicando como foram definidas e trabalhadas, como indicado na Tabela 7. Isso garante que a análise seja construída sobre uma base sólida e bem compreendida (Grandvalet, Bengio, 2004). Isso frequentemente envolve a visualização de dados, pois os gráficos podem revelar padrões, características e tendências que não são imediatamente óbvios em tabelas numéricas (Huang *et al.*, 2015).

O entendimento aprofundado dos dados vai além das estatísticas descritivas básicas, incluindo a identificação de padrões, a detecção de outliers e a avaliação da diversidade e cardinalidade dos dados em cada coluna. Esse processo permite reconhecer as principais variáveis de interesse, possíveis confundidores e variáveis de controle, orientando as análises subsequentes (Li, 2022). Para isso, foi desenvolvido um script apresentado no Quadro 34, utilizando a função `.nunique()` da biblioteca *Pandas*.

Quadro 34 – Script utilizado para demonstrar os valores únicos de cada coluna do DF3_trabalhos_BR

```
import pandas as pd
# Carregar o arquivo CSV
file_path = '/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF3.csv'
df = pd.read_csv(file_path, sep=";")
# Contar o número de valores únicos em cada coluna
df.nunique()
```

Fonte: Dados de pesquisa (2024).

A função `.nunique()` permite identificar rapidamente a quantidade de valores distintos em cada coluna, fornecendo uma visão clara da diversidade dos dados (McKinney, 2012). O resultado da função `.nunique()` é apresentado na Tabela 9, exibindo os nomes dos campos e a quantidade de valores únicos em cada coluna.

Tabela 9 – Resultado da execução da função `nunique()` da biblioteca *Pandas*

Índice	Campo	Valores únicos
1	ANO_BASE	19
2	CD_GRANDE_AREA_CONHECIMENTO	9
3	NM_GRANDE_AREA_CONHECIMENTO	9
4	CD_AREA_CONHECIMENTO	248
5	NM_AREA_CONHECIMENTO	248
6	NM_AREA_AVALIACAO	59
7	CD_PROGRAMA	5047
8	NM_PROGRAMA	2631
9	SG_ENTIDADE_ENSINO	628
10	NM_ENTIDADE_ENSINO	586
11	NM_REGIAO	5
12	SG_UF_IES	27
13	ID_ADD_PRODUCAO_INTELECTUAL	1232870
14	ID_PRODUCAO_INTELECTUAL	1232759

Índice	Campo	Valores únicos
15	NM_PRODUCAO	1230781
16	DS_PALAVRA_CHAVE	1216795
17	DS_RESUMO	1221320
18	ID_GRAU_ACADEMICO	4
19	NM_GRAU_ACADEMICO	8
20	NM_IDIOMA	67
21	ID_PESSOA_ORIENTADOR	154498
22	NM_ORIENTADOR	144843
23	ID_PESSOA_DISCENTE	869431
24	NM_DISCENTE	1094131

Fonte: Dados de pesquisa (2024).

A Tabela 9 revela a diversidade dos dados no DF3_trabalhos_BR, destacando a variação no número de valores únicos por coluna. Observa-se que algumas variáveis, como ANO_BASE e NM_REGIAO, possuem poucos valores distintos, enquanto outras, como ID_PRODUCAO_INTELECTUAL e NM_DISCENTE, apresentam alta cardinalidade. Essas informações são essenciais para entender a estrutura dos dados e definir estratégias de análise, especialmente no tratamento de variáveis categóricas e na detecção de possíveis redundâncias.

No Quadro 35, pode ser visto o script utilizado para computar a contagem de frequência dos valores em uma coluna específica, este script é denominado SCF (script de contagem de frequência) no decorrer da pesquisa. A contagem de frequência é uma técnica estatística que quantifica o número de ocorrências de cada valor distinto em um conjunto de dados (Thorpe; Favia, 2012).

Quadro 35 – Contagem de frequência de cada coluna do DF3_trabalhos_BR

```
import pandas as pd
# Carregar o arquivo CSV
file_path = '/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF3.csv'
df = pd.read_csv(file_path, sep=";")
# Contar o número de valores únicos em cada coluna
df.nunique()
```

Fonte: Dados de pesquisa (2024).

Ao aplicar esse script no DataFrame DF3_trabalhos_BR, ele revela a prevalência de cada valor único, facilitando a identificação de padrões e tendências predominantes em cada coluna. O resultado da aplicação do SCF para o campo ANO_BASE do DF3_trabalhos_BR está apresentado na Tabela 10.

Tabela 10 – Resultado da execução do SCF para o campo ANO_BASE do DF3_trabalhos_BR

Índice	Campo	Valores únicos
1	2004	34917
2	2005	39695
3	2006	41647
4	2007	42819
5	2008	46750
6	2009	50167
7	2010	50903
8	2011	55554
9	2012	61053
10	2013	67534
11	2014	70668
12	2015	75884
13	2016	80278
14	2017	85310
15	2018	90469
16	2019	94503
17	2020	80114
18	2021	82238
19	2022	82367

Fonte: Dados de pesquisa (2024).

As contagens apresentadas na Tabela 10 estão alinhadas com as Tabelas 01 e 02, validando a integridade dos dados. Essa consistência reforça a confiabilidade dos procedimentos de coleta e análise, garantindo que as conclusões sejam baseadas em informações consistentes e coerentes.

A coluna ANO_BASE indica o ano de conclusão de cada trabalho acadêmico, servindo como um marcador temporal para a produção científica. Essa informação é essencial para acompanhar tendências e mudanças na produção acadêmica ao longo do tempo. Conforme mostrado na Tabela 10, a pesquisa abrange 19 anos, de 2004 a 2022, com um aumento geral no número de trabalhos concluídos, exceto em 2020 e 2021, onde houve uma leve redução.

As colunas **CD_GRANDE_AREA_CONHECIMENTO** e **NM_GRANDE_AREA_CONHECIMENTO** desempenham um papel base na estruturação e categorização das produções acadêmicas. Elas representam, respectivamente, os códigos e os nomes das Grandes Áreas do Conhecimento, permitindo a análise da distribuição das pesquisas entre diferentes campos científicos. Segundo a CAPES (2023), a classificação das Áreas do Conhecimento tem uma função prática, proporcionando estruturação e sistematização das informações científicas e tecnológicas. Essa categorização não apenas organiza as pesquisas, mas também apoia a análise do desenvolvimento científico e da formação de pesquisadores. A Tabela 11 traz o mapeamento das informações referentes às Grandes Áreas do Conhecimento dos estudos de DF3_trabalhos_BR.

Tabela 11 – Resultado da execução do SCF para os campos código e nome da grande área no DF3_trabalhos_BR

Código grande área	Quantidade ids grande área	Nome grande área	Quantidade nome grande área	Diferença
70000000	203867	CIÊNCIAS HUMANAS	203865	2
40000001	192620	CIÊNCIAS DA SAÚDE	192619	1
60000007	172775	CIÊNCIAS SOCIAIS APLICADAS	172775	0
30000009	137891	ENGENHARIAS	137891	0
90000005	135496	MULTIDISCIPLINAR	135495	1
50000004	122472	CIÊNCIAS AGRÁRIAS	122472	0
10000003	105648	CIÊNCIAS EXATAS E DA TERRA	105648	0
20000006	85756	CIÊNCIAS BIOLÓGICAS	85756	0
80000002	76329	LINGUÍSTICA, LETRAS E ARTES	76328	1

Fonte: Dados de pesquisa (2024).

Ao analisar os dados da Tabela 11, observa-se uma discrepância entre os códigos e os respectivos nomes das Grandes Áreas do Conhecimento. Embora se esperasse uma correspondência exata, pequenas diferenças foram identificadas. Uma explicação para essa discrepância pode ser a presença de campos nulos na Tabela 8, sugerindo inconsistências na

correspondência entre os códigos e os nomes das áreas. Levando isso em consideração, torna-se essencial examinar a origem dessas inconsistências e promover as correções adequadas, de modo a garantir a confiabilidade das análises posteriores. Para uma verificação mais apurada, foi criado um *script*, conforme mostra o Quadro 36.

Quadro 36 – Script desenvolvido para encontrar divergências entre os códigos e nomes da grande área do conhecimento

```
import pandas as pd
# Carregando o arquivo DF3
caminho = '/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF3.csv'
df = pd.read_csv(caminho, sep=',')
# Definindo as combinações válidas
combinacoes_validas = {
    70000000: "CIÊNCIAS HUMANAS", 40000001: "CIÊNCIAS DA SAÚDE", 60000007: "CIÊNCIAS SOCIAIS APLICADAS",
    30000009: "ENGENHARIAS", 90000005: "MULTIDISCIPLINAR", 50000004: "CIÊNCIAS AGRÁRIAS", 10000003: "CIÊNCIAS
    EXATAS E DA TERRA", 20000006: "CIÊNCIAS BIOLÓGICAS", 80000002: "LINGÜÍSTICA, LETRAS E ARTES"
}
# Filtrando o DataFrame para pegar apenas os que NÃO estão nas combinações válidas
filtro = ~df[['CD_GRANDE_AREA_CONHECIMENTO', 'NM_GRANDE_AREA_CONHECIMENTO']].apply(tuple,
axis=1).isin(combinacoes_validas.items())
df_desafio = df[filtro]
# Salvando o DataFrame filtrado
caminho_salvo = '/content/drive/MyDrive/CAPES/datasets/dados_abertos/primeiro_desafio.csv'
df_desafio.to_csv(caminho_salvo, sep=',', index=False)
```

Fonte: Dados de pesquisa (2024).

Após a execução do script retratado no Quadro 36, observou-se o retorno de 21 linhas, onde estas se referem aos anos de 2017 (com 6 registros), 2018 (com 8 registros), 2019 (com 6 registros) e 2020 (com 1 registro). Surpreendentemente, todas essas linhas não possuíam o nome da Grande Área do Conhecimento preenchido, embora cinco delas tivessem o código correspondente especificado.

Como todas as linhas possuem o código do programa preenchido, a correção dos dados será feita acessando a Plataforma Sucupira, inserindo o código do programa, realizando a busca e capturando a informação da "Área Básica". Em seguida, é necessário verificar o código da Grande Área na tabela da CAPES. Embora esse período registre um volume maior de informações na CAPES, a existência dessa lacuna indica que, mesmo em sistemas estruturados, falhas podem ocorrer, reforçando a importância de uma análise criteriosa.

Em seguida, a análise se concentra nos campos `CD_AREA_CONHECIMENTO` e `NM_AREA_CONHECIMENTO`, que representam os códigos e nomes das Áreas de Conhecimento associadas aos Programas de Pós-Graduação. Segundo a CAPES (2023), as Áreas do Conhecimento são organizadas em quatro níveis hierárquicos, abrangendo nove grandes áreas e 49 áreas de avaliação. Dentro dessas, encontram-se as áreas básicas, definidas como conjuntos inter-relacionados de saberes, estruturadas para orientar ensino, pesquisa e aplicações práticas.

A Tabela 9 mostra que ambos os campos possuem 248 valores distintos, indicando coerência inicial. No entanto, ao aplicar o SCF, foi identificada uma discrepância: embora ambas as colunas retornem 248 registros, alguns valores não coincidem. Por exemplo, o código 40100006 está associado a 68.628 registros, mas seu correspondente nominal, "MEDICINA", aparece 68.627 vezes. Para corrigir essas inconsistências, é utilizado o *script* do Quadro 37, atualizando os campos filtro e combinacoes_validas={} com os valores do SCF.

A aplicação do *script* revelou 81 registros problemáticos, distribuídos pelos anos 2008 (33 registros), 2009 (24), 2017 (9), 2018 (8), 2019 (6) e 2020 (1). Os registros de 2008 e 2009 parecem estar concentrados no mesmo programa, sugerindo a possibilidade de correção em lote. Em anos mais recentes, além de valores nulos, quatro registros contêm códigos sem os respectivos nomes.

Embora o número de inconsistências tenha aumentado em relação à análise anterior, a correção neste caso pode ser mais simples, uma vez que muitas discrepâncias estão agrupadas. Segue-se a mesma estratégia anterior, assegurando a integridade e confiabilidade dos dados na pesquisa.

Em seguida, a análise se concentra no campo do Nome da Área de Avaliação à qual o Programa de Pós-Graduação está associado. Em 2023, a CAPES reconhece oficialmente 49 áreas de avaliação, mas a Tabela 9 indica 59 valores distintos, levantando duas possibilidades: houve mudanças nas áreas entre 2004 e 2023 ou algumas categorias foram registradas incorretamente no *dataset*. Alguns exemplos sugerem possíveis inconsistências, como as categorias "FILOSOFIA/TEOLOGIA: SUBCOMISSÃO FILOSOFIA" e "FILOSOFIA/TEOLOGIA: subcomissão TEOLOGIA", que podem indicar registros duplicados ou classificados de forma inadequada. Essa discrepância exige uma verificação para adequação dos dados.

Para esclarecer essa questão, é necessário comparar os dados do DF3_trabalhos_BR com as áreas de avaliação oficialmente reconhecidas pela CAPES em 2023. Para isso, foi aplicado o SCF, cujos resultados são apresentados na Tabela 12.

Tabela 12 – Áreas de avaliação encontradas no DataFrame com o uso SCF para o campo NM_AREA_AVALIACAO

Índice	Nome da área de avaliação	Quantidade de trabalhos
1	EDUCAÇÃO	78650
2	INTERDISCIPLINAR	71253

Índice	Nome da área de avaliação	Quantidade de trabalhos
3	CIÊNCIAS AGRÁRIAS I	65855
4	DIREITO	49711
5	MEDICINA I	45540
6	ENGENHARIAS III	43050
7	LETRAS / LINGUÍSTICA	36159
8	BIODIVERSIDADE	33961
9	ODONTOLOGIA	33575
10	ENGENHARIAS I	32834
11	ADMINISTRAÇÃO PÚBLICA E DE EMPRESAS, CIÊNCIAS CONTÁBEIS E TURISMO	32618
12	ENGENHARIAS IV	31819
13	ENGENHARIAS II	30512
14	PSICOLOGIA	30402
15	ENSINO	29300
16	QUÍMICA	27362
17	CIÊNCIA DA COMPUTAÇÃO	26258
18	HISTÓRIA	25162
19	SAÚDE COLETIVA	24567
20	CIÊNCIAS BIOLÓGICAS II	24412
21	MEDICINA VETERINÁRIA	23941
22	LINGUÍSTICA E LITERATURA	23064
23	ENFERMAGEM	21147
24	CIÊNCIAS BIOLÓGICAS I	20423
25	ZOOTECNIA / RECURSOS PESQUEIROS	20239
26	MEDICINA II	19588
27	ADMINISTRAÇÃO, CIÊNCIAS CONTÁBEIS E TURISMO	18867
28	EDUCAÇÃO FÍSICA	18476
29	ECONOMIA	18378
30	MATEMÁTICA / PROBABILIDADE E ESTATÍSTICA	18226
31	GEOGRAFIA	18187
32	SOCIOLOGIA	17907
33	ASTRONOMIA / FÍSICA	17731
34	CIÊNCIAS AMBIENTAIS	17698

Índice	Nome da área de avaliação	Quantidade de trabalhos
35	FARMÁCIA	15173
36	GEOCIÊNCIAS	14039
37	CIÊNCIA DE ALIMENTOS	13964
38	COMUNICAÇÃO E INFORMAÇÃO	13475
39	BIOTECNOLOGIA	10906
40	CIÊNCIAS BIOLÓGICAS III	10073
41	ARQUITETURA, URBANISMO E DESIGN	9914
42	ARTES / MÚSICA	9383
43	CIÊNCIA POLÍTICA E RELAÇÕES INTERNACIONAIS	8804
44	SERVIÇO SOCIAL	8759
45	PLANEJAMENTO URBANO E REGIONAL / DEMOGRAFIA	8343
46	MEDICINA III	8164
47	CIÊNCIAS SOCIAIS APLICADAS I	7062
48	ANTROPOLOGIA / ARQUEOLOGIA	6997
49	FILOSOFIA	6915
50	ARTES	6674
51	MATERIAIS	6127
52	NUTRIÇÃO	5903
53	ARQUITETURA E URBANISMO	4643
54	FILOSOFIA / TEOLOGIA:subcomissão FILOSOFIA	3805
55	FILOSOFIA/TEOLOGIA:subcomissão TEOLOGIA	2700
56	CIÊNCIAS DA RELIGIÃO E TEOLOGIA	2386
57	TEOLOGIA	1281
58	FILOSOFIA/TEOLOGIA:SUBCOMISSÃO TEOLOGIA	328
59	FILOSOFIA/TEOLOGIA:SUBCOMISSÃO FILOSOFIA	149

Fonte: Dados de pesquisa (2024).

É preciso também não negligenciar as entradas nulas para essa coluna no DataFrame: as 31 linhas. No entanto, como o código do programa está presente em todas as linhas, ele pode servir como um guia confiável para rastrear e corrigir possíveis inconsistências, conforme observado em etapas anteriores da pesquisa. A precisão no tratamento dos dados garante a validade e confiabilidade dos resultados.

A investigação agora se volta para os campos código e nome do programa de pós-graduação. O código do programa é gerado a partir do código da Instituição de Ensino Superior (IES) na CAPES, pois cada programa está vinculado a uma IES específica. Esse código é único e serve como identificador do programa, enquanto o nome do programa é fornecido pela instituição e aprovado pelo Ministério da Educação, funcionando como referência descritiva.

Uma análise na Tabela 9 indica uma discrepância: o campo código do programa apresenta 5047 valores distintos, enquanto o nome do programa tem apenas 2631 valores únicos. À primeira vista, isso sugere uma inconsistência. No entanto, ao aplicar o SCF, surge uma possível explicação: o nome do programa "EDUCAÇÃO" é o mais recorrente, com 62450 ocorrências, enquanto o código "31075010001P2" é o mais frequente, com 6664 registros. Isso sugere que, enquanto um mesmo nome de programa pode aparecer em diferentes IES, cada instituição tem um código único para seu programa, resultando em múltiplos códigos associados a um mesmo nome.

Para validar essa hipótese, será necessário consultar a Plataforma Sucupira, onde é possível baixar a lista oficial de programas e comparar sistematicamente os códigos e nomes, utilizando funções específicas para esse fim. A ausência de valores nulos nesses campos simplifica a análise e reduz a necessidade de limpeza adicional, reforçando a importância de uma coleta de dados estruturada para garantir a integridade das informações na pesquisa.

A análise agora se volta para os campos Sigla e Nome da IES associada a cada Tese ou Dissertação. A Tabela 9 revela uma discrepância: enquanto o campo de sigla contém 628 valores distintos, o campo nome apresenta 686, uma diferença de 58 registros. Como ambos os campos devem representar a mesma instituição, essa variação sugere inconsistências. Além disso, a ausência de um identificador único (como um código de ID) dificulta a padronização e a eliminação de ambiguidades.

A aplicação do SCF confirma essa inconsistência. Por exemplo, a sigla FGV/SP está associada a 3898 registros, mas aparece vinculada a dois nomes diferentes: "FUNDAÇÃO GETÚLIO VARGAS (SP)" (2050 registros) e "FUNDAÇÃO GETÚLIO VARGAS/SP" (2201 registros). A soma desses valores coincide com o total associado à sigla, evidenciando variações na nomenclatura.

Como esses campos são textuais e não possuem um código identificador único, a solução mais eficiente é comparar esses dados com as informações estruturadas disponíveis no portal da CAPES, que contém o nome do programa, seu código e as siglas e nomes das

IES. Essa estrutura tabular facilita a validação e padronização por meio de códigos de programação, contribuindo para a precisão da pesquisa.

A análise segue com a região geográfica da IES, um campo relevante para entender a distribuição dos programas de pós-graduação no Brasil. A Tabela 9 indica a presença de cinco valores distintos, o que corresponde às cinco regiões do país. Essa correspondência sugere coerência nos dados, cujos detalhes e contagens estão apresentados na Tabela 13, resultado da aplicação do SCF.

Tabela 13 – Dados dos trabalhos encontrados por região no DF3_trabalhos_BR

Índice	Região	Total de trabalhos encontrados
1	SUDESTE	638305
2	SUL	250729
3	NORDESTE	209044
4	CENTRO-OESTE	86692
5	NORTE	48089
TOTAL		1232859

Fonte: Dados de pesquisa (2024).

A soma dos registros por região totaliza 1.232.859, apresentando uma discrepância de 11 registros em relação ao total esperado. Esses registros não possuem informações na coluna de Região. Para corrigir essa lacuna, pode-se adotar a mesma abordagem utilizada nos campos Sigla e Nome da IES, utilizando a planilha da CAPES, que detalha a localização estadual de cada programa. Com essas informações, um *script* pode preencher automaticamente as regiões ausentes, garantindo a integridade do DF3_trabalhos_BR.

A análise avança para o campo da localização estadual das IES, um dado que ajuda a entender a distribuição geográfica dos programas de pós-graduação no Brasil. Junto com a informação sobre regiões, essa variável pode oferecer uma visão detalhada do acesso à pós-graduação no país. Uma visualização geográfica, como um mapa de calor, pode destacar áreas de maior e menor concentração de programas, fornecendo insights sobre desigualdades na distribuição da educação de pós-graduação. A Tabela 9 mostra 27 valores distintos, o que corresponde exatamente aos 26 estados e ao Distrito Federal, sugerindo consistência nos dados.

A Tabela 14 apresenta os resultados da aplicação do SCF, detalhando a distribuição dos registros por estado.

Tabela 14 – Dados dos trabalhos encontrados por UF no DF3_trabalhos_BR

Índice	Estado	Total de trabalhos encontrados
1	SP	338330
2	RJ	155355
3	MG	126669
4	RS	116685
5	PR	84341
6	SC	49702
7	PE	45645
8	BA	41437
9	DF	36828
10	CE	35450
11	PB	28701
12	PA	24991
13	RN	24887
14	GO	24404
15	ES	17952
16	MS	14295
17	AM	12279
18	MT	11165
19	SE	9985
20	MA	7932
21	PI	7732
22	AL	7275
23	TO	3797
24	RO	2535
25	AC	1785
26	RR	1509
27	AP	1193
TOTAL		1232859

Fonte: Dados de pesquisa (2024).

Uma análise mais aprofundada revela uma leve discrepância: o total de registros não corresponde exatamente ao número de 1.232.859 trabalhos. Esta diferença se deve à existência de 11 registros com valores ausentes para a coluna de estados. Este cenário remete

à necessidade de adotar uma abordagem similar à empregada para o campo "nome da região". A planilha disponibilizada pela Capes será fundamental para guiar a correção e o preenchimento dessas lacunas, assegurando a integridade e precisão do DataFrame em análise.

A análise avança para o campo `ID_ADD_PRODUCAO_INTELECTUAL`, que serve como identificador da Tese ou Dissertação no Ano Base e no Programa na Base de Dados da CAPES. A falta de informações detalhadas da CAPES sobre este campo é notável. No entanto, conforme evidenciado na Tabela 9, é possível discernir que cada registro possui um identificador único, totalizando 1.232.870 entradas distintas. Em sequência, o foco se direciona para o campo `ID_PRODUCAO_INTELECTUAL`, que representa o Identificador da Tese ou Dissertação na própria Base de Dados da CAPES. Semelhante ao campo anterior, informações adicionais sobre sua natureza são escassas. Entretanto, uma observação cuidadosa na Tabela 9 revela que existem 1.232.759 registros para este campo, resultando em uma diferença de 111 entradas em comparação com o campo anterior.

Para investigar as discrepâncias nos identificadores das teses e dissertações, foi desenvolvido um script em *Python*, apresentado no Quadro 37. Esse *script* permite filtrar e visualizar registros específicos, seja por um valor único em determinada coluna ou pela presença de valores nulos.

Quadro 37 – Script desenvolvido para visualizar valores únicos por coluna e linha específicos

```
import pandas as pd
# Carregando o arquivo DF3
caminho = '/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF3.csv'
df = pd.read_csv(caminho, sep=',')
def filter_and_display(COLUNA, VALOR):
    # Filtrando o DataFrame
    if VALOR in [None, "null", "NULL", "", " "]:
        filtered_df = df[df[COLUNA].isna()]
    else:
        filtered_df = df[df[COLUNA] == VALOR]
    # Exibindo o DataFrame filtrado
    return filtered_df
# Definindo a coluna e o valor desejado
COLUNA = 'ID_PRODUCAO_INTELECTUAL'
VALOR = 10915599
# Executando a função e exibindo os dados
display_data = filter_and_display(COLUNA, VALOR)
display_data
```

Fonte: Dados de pesquisa (2024).

A aplicação desse *script* possibilitou uma análise detalhada das diferenças entre os campos `ID_PRODUCAO_INTELECTUAL` e `ID_ADD_PRODUCAO_INTELECTUAL`. Por exemplo, ao buscar o valor 10915599 no campo `ID_PRODUCAO_INTELECTUAL`, identificou-se sua associação com dois registros distintos, um em 2020 e outro em 2021,

ambos vinculados ao programa DESENVOLVIMENTO LOCAL da UNIVERSIDADE CATÓLICA DOM BOSCO (UCDB).

Essa observação sugere que as 111 discrepâncias podem estar relacionadas à duplicidade de registros. Para resolver essa questão, uma abordagem viável é a consulta direta ao programa de pós-graduação ou ao *DataFrame* de discentes, a fim de identificar o ano exato da defesa de cada trabalho, garantindo a precisão e integridade dos dados.

A análise segue para o campo título da produção, que deveria conter apenas entradas únicas, mas a Tabela 9 aponta 2.089 registros duplicados. Para aprofundar essa questão, foi desenvolvido um novo *script*, apresentado no Quadro 38, que identifica valores repetidos em NM_PRODUCAO.

Quadro 38 – Script desenvolvido para visualizar valores únicos com resultado maior que 1

```
import pandas as pd
# Carregando o arquivo DF3
caminho = '/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF3.csv'
df = pd.read_csv(caminho, sep=',')
def filter_values_with_more_than_two_records(column_name, DataFrame=df):
    # Contando os valores únicos
    value_counts = DataFrame[column_name].value_counts()
    # Filtrando valores que possuem mais de 2 registros
    filtered_values = value_counts[value_counts > 1]
    return filtered_values
# Exemplo de uso:
column_to_check = 'NM_PRODUCAO'
results = filter_values_with_more_than_two_records(column_to_check)
results
```

Fonte: Dados de pesquisa (2024).

A execução desse *script* revelou padrões preocupantes, como a repetição do título "O ALUNO DO PROGRAMA DE MESTRADO DO IMPA TEM A OPÇÃO DE NÃO DEFENDER DISSERTAÇÃO, O ALUNO PODE CURSAR DUAS DISCIPLINAS DE DOUTORADO", que aparece 145 vezes. Além disso, registros genéricos ou incompletos, como "x" (19 ocorrências), "..." (6 ocorrências) e "Pendente" (3 ocorrências), exigem verificação para determinar sua validade e impacto nos resultados.

Diante desse cenário, a padronização dos dados é essencial. Registros como "x", "." e "..." podem ser substituídos por um marcador "não informado", enquanto casos relacionados à ausência de defesa podem ser unificados sob uma nomenclatura padronizada, como "Não houve defesa de Mestrado". Essas medidas são fundamentais para garantir a coerência e integridade do DF3_trabalhos_BR.

A investigação agora se concentra no campo DS_PALAVRA_CHAVE, responsável por catalogar as palavras-chave das teses e dissertações. Inicialmente, esperava-se que cada linha contivesse uma entrada única, totalizando 1.232.870 registros distintos. No entanto, a

Tabela 9 revela 1.216.795 registros preenchidos, resultando em uma discrepância de mais de 16.000 linhas.

A aplicação do *script* do Quadro 16, ajustado para a coluna DS_PALAVRA_CHAVE, evidenciou padrões incomuns, como placeholders e valores genéricos. Os mais recorrentes incluem "." (1.213 linhas), "X" (754 linhas), "XXX" (465 linhas), "? " (461 linhas), "XX " (437 linhas), "não informou" (413 linhas), "- " (247 linhas), "... " (229 linhas), "EM BREVE" (210 linhas), "AGUARDANDO" (210 linhas), "PENDENTE VERSÃO FINAL" (210 linhas), "x" (194 linhas), "1" (185 linhas), "A " (184 linhas), "xx" (153 linhas), "FALTA" (132 linhas), "XXXX" (117 linhas), "PALAVRAS-CHAVE" (98 linhas), "A INFORMAR" (97 linhas), "a " (95 linhas), "NÃO INFORMADO" (94 linhas), ".." (94 linhas), "PENDENTE" (86 linhas), "xxx " (82 linhas), "A DEFINIR" (61 linhas), "não informado" (61 linhas), "EM ANDAMENTO" (58 linhas), "BREVE" (55 linhas), "NÃO DISPONÍVEL" (53 linhas), "AGUARDANDO VERSÃO " (52 linhas), entre outros. Além disso, a primeira palavra-chave válida, "geografia", aparece 77 vezes, sugerindo que alguns trabalhos podem conter apenas um termo, o que compromete a representatividade do campo.

Outro problema identificado é a falta de padronização na inserção das palavras-chave. Cada instituição utiliza um formato diferente, separando os termos por vírgula (,), ponto e vírgula (;), traço (-), ponto (.), underline (_), números ou espaços em branco. Essa inconsistência dificulta análises como a identificação das palavras mais utilizadas por área do conhecimento ou por ano.

A revisão desse campo exige uma abordagem criteriosa para validar, padronizar e, se necessário, corrigir as entradas. A interpretação precisa desses dados é fundamental para garantir a integridade do DataFrame e a robustez da pesquisa.

A análise agora se concentra no campo DS_RESUMO, responsável pelo armazenamento dos resumos das produções acadêmicas. Esperava-se que esse campo contivesse 1.232.870 registros únicos, porém, a Tabela 9 aponta 1.221.320 entradas preenchidas, indicando uma diferença de mais de 11.000 registros.

A execução do *script* do Quadro 16 revelou padrões semelhantes aos observados em outros campos, com a presença de entradas genéricas ou inadequadas. Entre os registros mais frequentes estão "." (947 linhas), "X" (499 linhas), "x" (417 linhas), "XXX" (335 linhas), "XX" (299 linhas), "... " (245 linhas), "- " (241 linhas), "RESUMO" (236 linhas), "PENDENTE VERSÃO FINAL;" (208 linhas), "EM BREVE" (207 linhas), "não informou" (195 linhas), "A" (161 linhas), "AGUARDANDO" (159 linhas), "xxx" (157 linhas), "xx" (135 linhas), "XXXX" (123 linhas), "1" (97 linhas), "FALTA" (95 linhas), "a" (75 linhas), "XXXXXX" (75

linhas), "." (74 linhas), "EM ANDAMENTO" (70 linhas), "NÃO DISPONÍVEL" (68 linhas), "Resumo" (67 linhas), "xxxx" (66 linhas), "A INFORMAR" (56 linhas), "AGUARDANDO VERSÃO FINAL" (54 linhas), "NÃO INFORMADO" (53 linhas), entre outros.

Além dessas, um texto aparece 66 vezes, indicando possível uso indevido do campo: "Dentro do processo altamente seletivo do IMPA e visando criar uma maior unidade entre o Mestrado e o Doutorado, declaramos, para os devidos fins, que a Comissão de Ensino decidiu que para a obtenção do grau de Mestre, os alunos poderão substituir a dissertação de Mestrado por duas disciplinas de Doutorado, conforme as normas do Catálogo de Pós-Graduação do IMPA". Isso sugere que, em alguns casos, marcadores ou notas explicativas foram inseridos no lugar dos resumos reais. Diante desse cenário, será necessária uma revisão detalhada para diferenciar entradas legítimas de *placeholders*, garantindo a consistência e confiabilidade dos dados.

A análise dos campos NM_PRODUCAO, DS_PALAVRA_CHAVE e DS_RESUMO indica a necessidade de validações mais rigorosas por parte da CAPES. Entradas simplificadas como ".", "x" e "?" não deveriam ser aceitas sem uma verificação prévia. Definir um comprimento mínimo para os campos e estabelecer textos padronizados para programas com normativas específicas poderiam melhorar a qualidade dos dados. Além disso, a inclusão de campos adicionais para indicar a ausência de defesa e um campo para justificativa poderia aprimorar a estrutura das informações disponibilizadas. Fica evidente que é necessário maior rigor na inserção dos dados nas plataformas da CAPES pelos programas de pós-graduação, tendo em vista garantir que as informações representem fielmente os trabalhos acadêmicos. A precisão dos registros é necessária para que as análises e os resultados gerados sejam baseados em dados confiáveis.

Seguindo em frente, foi analisado o campo ID_GRAU_ACADEMICO, com foco em mestrado e doutorado. A análise da estrutura dos graus acadêmicos na pós-graduação revelou maior variação do que o esperado. Inicialmente, identificaram-se dois graus principais (mestrado e doutorado), mas a Tabela 9 mostrou a existência de quatro códigos distintos para o identificador de grau acadêmico e oito denominações diferentes para os nomes dos graus, indicando inconsistências na categorização, conforme mostra a Tabela 15.

Tabela 15 – Nomes dos graus acadêmicos encontrados no DF3_trabalhos_BR

Índice	Nome do grau ou modalidade	Total de trabalhos encontrados
1	MESTRADO	486369

Índice	Nome do grau ou modalidade	Total de trabalhos encontrados
2	Mestrado	301641
3	DOUTORADO	206549
4	MESTRADO PROFISSIONAL	116321
5	Doutorado	96022
6	Profissionalizante	25841
7	DOUTORADO PROFISSIONAL	120
8	Mestrado e Profissionalizante	1
TOTAL		1232864

Fonte: Dados de pesquisa (2024).

A Tabela 15 apresenta os nomes dos graus acadêmicos encontrados no DF3_trabalhos_BR, evidenciando variações na nomenclatura e uma discrepância de seis registros sem especificação do grau acadêmico. Essa inconsistência reforça a necessidade de padronização dos nomes dos graus, alinhando-os ao campo ID_GRAU_ACADEMICO para garantir a precisão das análises. A uniformização desses dados permitirá uma avaliação mais clara do grau defendido em cada trabalho acadêmico e contribuirá para a consistência e confiabilidade da pesquisa, além de facilitar a interpretação da estrutura dos programas de pós-graduação.

A análise do campo IDIOMA, conforme a Tabela 9, identificou 67 valores únicos e 37 registros nulos. A aplicação do SCF revelou variações na forma de registro, como "PORTUGUES" (787.604 ocorrências) e "Português" (416.716 ocorrências), além de diferenças semelhantes em outros idiomas, como "INGLES" (20.570), "FRANCES" (113) e "Francês" (68). Além disso, há descrições genéricas, como "Outro" (4.763 ocorrências), "Diversos" (34), "Português/Inglês" (4) e "Português/resumo, Inglês/texto" (1), que dificultam a categorização. Para padronização, recomenda-se a unificação das variações sob uma nomenclatura única por meio de um *script* e a substituição de descrições ambíguas por "Não Informado" quando não houver uma classificação clara. A padronização desse campo é também fundamental para a análise da distribuição linguística dos trabalhos de pós-graduação no Brasil, e podem reduzir inconsistências e melhorar a qualidade e confiabilidade dos dados na plataforma da CAPES.

A análise dos campos ID_PESSOA_ORIENTADOR e NM_ORIENTADOR examina a identificação dos orientadores na base da CAPES e sua correspondência com o cadastro de Pessoa Física da Receita Federal. O identificador é único e não varia conforme o vínculo

acadêmico ou programa de pós-graduação. Inicialmente, esperava-se que ambas as colunas contivessem o mesmo número de registros. No entanto, observou-se que, de 2004 a 2012, o campo ID_PESSOA_ORIENTADOR não estava disponível. Para preencher essa lacuna, utilizou-se o campo DocumentoOrientador_1, que contém um identificador denominado "CPF Tarjado" pela CAPES, permitindo que um mesmo ID fosse associado a diferentes professores.

A Tabela 9 mostra 154.498 registros únicos para ID_PESSOA_ORIENTADOR e 144.843 nomes únicos em NM_ORIENTADOR, indicando uma discrepância de aproximadamente 10.000 registros. Para investigar essa diferença, foi aplicado o *script* do Quadro 39, que agrupou os orientadores por seus respectivos identificadores.

Quadro 39 – Script desenvolvido para visualizar identificadores dos Professores de 2004 a 2012

```
import pandas as pd
# Carregando o arquivo DF3
caminho = '/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF3.csv'
df = pd.read_csv(caminho, sep=';')
def filter_values_with_more_than_two_records(column_name, DataFrame=df):
    # Contando os valores únicos
    value_counts = DataFrame[column_name].value_counts()
    # Filtrando valores que possuem mais de 2 registros
    filtered_values = value_counts[value_counts > 2]
    return filtered_values
# Exemplo de uso:
column_to_check = 'NM_ORIENTADOR'
results = filter_values_with_more_than_two_records(column_to_check)
results
```

Fonte: Dados de pesquisa (2024).

A aplicação do *script* revelou um erro na estrutura dos dados: espaços seguidos de parênteses no início de alguns identificadores, o que resultou na criação de novos identificadores duplicados. A correção dessas inconsistências garante que a identificação dos orientadores seja precisa e que os registros reflitam corretamente a relação entre os docentes e os trabalhos orientados.

A necessidade de um *script* adicional surgiu, conforme apresentado no Quadro 40, após a exportação dos dados para .CSV e sua visualização no LibreOffice Calc, onde foram identificadas variações na escrita dos nomes dos orientadores. Exemplos incluem "ABDIAS MAGALHAES GOMES" e "ABDIAS MAGALHÃES GOMES", diferenciados pelo uso do til, e "ABEL WASHINGTON DE ALBUQUERQUE" versus "ABEL WASHINGTON DE ALBUQUERQUE", com um "H" extra. Além disso, observou-se que um mesmo identificador estava associado a diferentes nomes, contrariando a premissa de que cada identificador deve ser único. Essas inconsistências indicam a necessidade de padronização dos nomes e

verificação da correspondência entre identificadores e orientadores, garantindo a coerência dos registros.

Quadro 40 – Script desenvolvido para visualizar nomes dos orientadores duplicados

```
import pandas as pd
# Carregando o arquivo DF3
caminho = '/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF3.csv'
df = pd.read_csv(caminho, sep=';')
df['ID_PESSOA_ORIENTADOR'] = df['ID_PESSOA_ORIENTADOR'].astype(str)
# Agrupar por NM_ORIENTADOR e concatenar IDs únicos de PESSOA_ORIENTADOR associados, separados por '-'
df_agrupado = df.groupby('NM_ORIENTADOR')['ID_PESSOA_ORIENTADOR'].apply(lambda x: '-'.join(x.unique())).reset_index()
# Ordenar o DataFrame pelo nome do orientador
df_agrupado.sort_values(by='NM_ORIENTADOR', inplace=True)
# Salvar o resultado em um arquivo CSV, usando ';' como separador
caminho_saida = '/content/drive/MyDrive/CAPES/datasets/nome_professores.csv'
df_agrupado.to_csv(caminho_saida, index=False, sep=';')
print(f'Arquivo salvo com sucesso em: {caminho_saida}')
```

Fonte: Dados de pesquisa (2024).

A execução do script para análise dos campos de identificação e nomes dos orientadores revelou inconsistências e possíveis soluções. Algumas linhas continham múltiplos nomes separados por vírgulas, indicando a participação dos professores em bancas de defesa. A correção dessas discrepâncias envolve normalizar os nomes e ajustar os identificadores. Esse último processo exige uma verificação manual, incluindo a exportação dos dados agrupados para um arquivo .CSV e sua análise em uma planilha eletrônica. Essa etapa possibilita a apresentação de diferentes panoramas sobre os orientadores, permitindo análises por programa, universidade, região, estado e cidade, ampliando as possibilidades de interpretação dos dados.

Prosseguindo, a análise dos campos ID_PESSOA_DISCENTE e NM_DISCENTE examina a identificação dos discentes na base de dados da CAPES e sua correspondência com o cadastro de Pessoa Física da Receita Federal. Esse identificador é único e não se altera com a mudança de vínculo acadêmico, grau ou programa de pós-graduação.

Esperava-se que ambos os campos apresentassem o mesmo número de registros únicos, mas essa correspondência não se confirmou, o que sugeriu a existência de inconsistências. Para verificar e corrigir essas diferenças, é realizada uma consulta ao conjunto de dados "Discentes da Pós-Graduação *Stricto Sensu* do Brasil", buscando assegurar a precisão das informações no DF3_trabalhos_BR. A análise do DF3_trabalhos_BR também evidenciou a necessidade de integrar novos conjuntos de dados para complementar e validar as informações. A inclusão da planilha de quantitativo de programas e do relatório de situação

dos programas é importante, pois contêm dados como código do programa, nome do programa, sigla e nome da IES, UF, situação em 2023 e modalidade do programa.

Em síntese, o DF3_trabalhos_BR apresenta inconsistências que exigem ajustes, como registros atípicos, variações nos códigos e erros de registro. A incorporação de dados adicionais sobre docentes, discentes e programas de pós-graduação pode ser uma solução para tratar essas falhas. Para garantir análises precisas e confiáveis, é necessário compreender e integrar esses conjuntos de dados de forma estruturada, permitindo a obtenção de informações mais consistentes e relevantes para a pesquisa.

Com a conclusão da Fase 3, de Entendimento dos Dados, inicia-se a Fase 4 - Preparação dos Dados, onde os ajustes necessários serão implementados para garantir a integridade e confiabilidade do DF3_trabalhos_BR. A análise nesta Fase 3 revelou inconsistências que precisam ser corrigidas antes da aplicação de métodos analíticos.

4.5.1.5.4 Fase 4 - Preparação dos dados

Nesta fase, é feita a preparação dos dados para análise. Isso pode envolver a transformação de variáveis, a criação de novas variáveis derivadas e a preparação de subconjuntos de dados específicos para análises focadas (Witten *et al.*, 2005). Uma das principais atividades é a conversão de planilhas em *DataFrames*, devido ao volume de programas e relatórios sobre a situação dos cursos. Os documentos, originalmente estruturados em planilhas, devem ser integrados ao ambiente analítico para possibilitar a manipulação eficiente dos dados e permitir uma visão consolidada (Nanda *et al.*, 2000; McKinney, 2012).

Os conjuntos de dados referentes a Programas, Docentes e Discentes da Pós-Graduação precisam ser convertidos e integrados ao DF3_trabalhos_BR. Antes dessa incorporação, é preciso compreender a estrutura e o conteúdo de cada planilha, evitando distorções que possam comprometer a qualidade da pesquisa (McKinney, 2012). A preparação dos dados segue uma abordagem sistemática, garantindo consistência analítica e replicabilidade dos resultados, em três passos: análise e ajustes do conjunto de dados dos Programas, entendimento e ajustes do conjunto de dados dos Docentes e entendimento e ajustes do conjunto de dados dos Discentes, descritos a seguir, em três passos distintos, de entendimento e ajustes dos conjuntos de dados.

Passo 1: Entendimento e ajustes do conjunto de dados dos Programas

Este passo tem como objetivo compreender os programas de pós-graduação *stricto sensu* no Brasil, etapa para a análise e formulação das conclusões. A CAPES disponibiliza quatro conjuntos de dados sobre esses programas, abrangendo o período de 2013 a 2022. Esses dados incluem nota do programa, área de conhecimento, ano de início e informações da IES vinculada. Os arquivos de 2013 a 2016 não contêm o campo Nome da Área Básica, enquanto os de 2017 a 2022 possuem essa informação, permitindo a consolidação dos dados. A seguir apresentam-se a análise do conjunto de dados (item 1) e dos ajustes de dados (item 2).

Análise do conjunto de dados dos Programas

A análise do conjunto de dados identificou campos como Nome da Grande Área do Conhecimento, Nome da Área do Conhecimento, Nome da Subárea do Conhecimento e Nome da Especialidade, que não possuem códigos correspondentes. Para suprir essa ausência, os códigos foram obtidos e integrados por meio de programação, exigindo a criação do *DataFrame* DF4_programas_BR. O Quadro 41 apresenta o processo de criação do DF4_programas_BR, detalhando os ajustes realizados para garantir a estruturação correta dos dados.

Quadro 41 – Script desenvolvido para criação do DF4_programas_BR

```
import pandas as pd
# Definindo as colunas do DataFrame
colunas = [
    "AN_BASE", "CD_GRANDE_AREA_CONHECIMENTO", "NM_GRANDE_AREA_CONHECIMENTO",
    "CD_AREA_CONHECIMENTO", "NM_AREA_CONHECIMENTO", "NM_SUBAREA_CONHECIMENTO",
    "NM_ESPECIALIDADE", "CD_AREA_AVALIACAO", "NM_AREA_AVALIACAO", "CD_ENTIDADE_CAPES",
    "CD_ENTIDADE_EMEC", "SG_ENTIDADE_ENSINO", "NM_ENTIDADE_ENSINO", "CS_STATUS_JURIDICO",
    "DS_DEPENDENCIA_ADMINISTRATIVA", "DS_ORGANIZACAO_ACADEMICA", "NM_REGIAO", "SG_UF_PROGRAMA",
    "NM_MUNICIPIO_PROGRAMA_IES", "NM_MODALIDADE_PROGRAMA", "CD_PROGRAMA_IES", "NM_PROGRAMA_IES",
    "NM_PROGRAMA_IDIOMA", "NM_GRAU_PROGRAMA", "CD_CONCEITO_PROGRAMA", "AN_INICIO_PROGRAMA",
    "AN_INICIO_CURSO", "DS_SITUACAO_PROGRAMA"
]
# Criando um DataFrame vazio com as colunas definidas
df = pd.DataFrame(columns=colunas)
# Salvando o DataFrame em um arquivo CSV
caminho = "/content/drive/MyDrive/CAPES/datasets/DF4.csv"
df.to_csv(caminho, sep=";", encoding='UTF-8', index=False)
```

Fonte: Dados de pesquisa (2024).

Os conjuntos de dados da CAPES já possuem os campos preenchidos, mas, para garantir uniformidade e compatibilidade, foi necessário ajustar o *encoding* para UTF-8 e padronizar o separador como ";". O Quadro 42 apresenta o *script* utilizado para essa conversão.

Quadro 42 – Script desenvolvido para alteração do encoding e separador dos arquivos

```
import os
import pandas as pd
# Definindo o caminho da pasta
caminho_pasta = "/content/drive/MyDrive/CAPES/DF4/"
# Listando todos os arquivos .csv na pasta
arquivos_csv = [f for f in os.listdir(caminho_pasta) if f.endswith('.csv')]
# Iterando sobre cada arquivo .csv
for arquivo in arquivos_csv:
    try:
        # Tentando ler o arquivo com encoding 'utf-8' e detecção automática do separador
        df = pd.read_csv(os.path.join(caminho_pasta, arquivo), encoding='utf-8', sep=None, engine='python')
    except UnicodeDecodeError:
        # Se ocorrer um erro de decodificação, tenta ler com encoding 'ISO-8859-1'
        df = pd.read_csv(os.path.join(caminho_pasta, arquivo), encoding='ISO-8859-1', sep=None, engine='python')
    except pd.errors.ParserError:
        # Se ocorrer um erro de análise, tenta pular linhas problemáticas
        df = pd.read_csv(os.path.join(caminho_pasta, arquivo), encoding='utf-8', error_bad_lines=False)
    # Salvando o arquivo com as especificações desejadas
    df.to_csv(os.path.join(caminho_pasta, arquivo), sep=";", encoding='UTF-8', index=False)
    print(f"Arquivo {arquivo} processado e salvo com sucesso!")
print("Todos os arquivos foram processados!")
```

Fonte: Dados de pesquisa (2024).

Com os arquivos preparados, a próxima etapa é a leitura e integração ao DF4_programas_BR, permitindo análises mais detalhadas. O Quadro 43 apresenta o processo de consolidação dos arquivos, que abrangem os dados dos programas de pós-graduação *stricto sensu* de 2013 a 2021.

Quadro 43 – Script desenvolvido juntar os dados dos arquivos do conjunto dos programas

```
import pandas as pd
# Caminho base para os arquivos
caminho_base = "/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF4/"
# Iterando sobre os anos de 2013 a 2021
for ano in range(2013, 2022):
    caminho_arquivo = f"{caminho_base}{ano}.csv"
    # Lendo o arquivo apenas para contar as linhas, sem carregar todo o conteúdo
    num_linhas = sum(1 for linha in open(caminho_arquivo))
```

Fonte: Dados de pesquisa (2024).

A metodologia adotada para a análise e estruturação do DF3_trabalhos_BR baseia-se no script do Quadro 43. A Tabela 16 apresenta as dimensões do DataFrame, a distribuição de valores por coluna e a tipologia de cada campo, garantindo a organização e consistência dos dados para a investigação.

Tabela 16 – Resultado do script demonstrado no Quadro 43

Índice	Nome do campo	Quantidade linhas	Tipo do campo	Valores únicos
1	AN_BASE	42587	int64	10
2	CD_GRANDE_AREA_CONHECIMENTO	0	float64	0

Índice	Nome do campo	Quantidade linhas	Tipo do campo	Valores únicos
3	NM_GRANDE_AREA_CONHECIMENTO	42587	object	9
4	CD_AREA_CONHECIMENTO	0	float64	0
5	NM_AREA_CONHECIMENTO	42587	object	81
6	NM_SUBAREA_CONHECIMENTO	42587	object	161
7	NM_ESPECIALIDADE	42587	object	90
8	CD_AREA_AVALIACAO	42587	int64	49
9	NM_AREA_AVALIACAO	42587	object	57
10	CD_ENTIDADE_CAPES	42587	int64	502
11	CD_ENTIDADE_EMEC	42587	object	366
12	SG_ENTIDADE_ENSINO	42587	object	568
13	NM_ENTIDADE_ENSINO	42587	object	631
14	CS_STATUS_JURIDICO	42587	object	4
15	DS_DEPENDENCIA_ADMINISTRATIVA	42587	object	2
16	DS_ORGANIZACAO_ACADEMICA	42587	object	6
17	NM_REGIAO	42587	object	5
18	SG_UF_PROGRAMA	42587	object	27
19	NM_MUNICIPIO_PROGRAMA_IES	42587	object	328
20	NM_MODALIDADE_PROGRAMA	42587	object	2
21	CD_PROGRAMA_IES	42587	object	4911
22	NM_PROGRAMA_IES	42587	object	2249
23	NM_PROGRAMA_IDIOMA	42587	object	3911
24	NM_GRAU_PROGRAMA	42587	object	6
25	CD_CONCEITO_PROGRAMA	42587	object	8
26	AN_INICIO_PROGRAMA	42587	int64	64
27	AN_INICIO_CURSO	42587	object	888
28	DS_SITUACAO_PROGRAMA	42587	object	2

Fonte: Dados de pesquisa (2024).

A análise da Tabela 16 mostra que a quantidade de valores por campo é uniforme, o que contribui para a integridade da análise. No entanto, a tipologia de alguns campos

apresenta discrepâncias em relação às expectativas iniciais. O campo AN_INICIO_CURSO, classificado como "object", sugere a presença de caracteres alfanuméricos em vez de datas ou números, divergindo da interpretação esperada. Da mesma forma, o campo CD_PROGRAMA_IES, que deveria conter um código numérico, também aparece como "object", indicando a necessidade de conversão para um formato adequado à sua função semântica. Esses ajustes são necessários para garantir a coerência dos dados e a precisão das análises subsequentes.

As contagens de valores entre campos correlatos apresentam diferenças, como nos casos de CD_AREA_AVALIACAO (49 registros) e NM_AREA_AVALIACAO (57 registros). Discrepâncias semelhantes ocorrem nos pares SG_ENTIDADE_ENSINO e NM_ENTIDADE_ENSINO e CD_PROGRAMA_IES e NM_PROGRAMA_IES, indicando variações na representação ou na completude dos dados. Para corrigir essas inconsistências, é necessário um processo de limpeza e padronização, incluindo a conversão dos tipos de campos para formatos compatíveis com as informações representadas.

A inspeção visual no LibreOffice Calc permitiu identificar campos vazios e verificar a codificação do arquivo. A análise indicou a necessidade de organizar os dados pelo nome da Instituição de Ensino. Também foram examinados os códigos institucionais, sendo que CD_ENTIDADE_CAPES está presente em todas as instituições, enquanto CD_ENTIDADE_EMEC apresentou inconsistências, tornando-se inadequado para uso. No campo NM_GRAU_PROGRAMA, programas que abrangem mestrado e doutorado são registrados como "MESTRADO/DOCTORADO". Para padronização e facilitar consultas futuras, essa nomenclatura foi convertida para uma representação numérica, conforme apresentado na Tabela 17.

Tabela 17 – Nomes dos graus acadêmicos encontrados no DF3_trabalhos_BR

De	Para
MESTRADO	1
MESTRADO/DOCTORADO	2
DOCTORADO	3
MESTRADO PROFISSIONAL	4
DOCTORADO PROFISSIONAL	5
MESTRADO PROFISSIONAL/ DOCTORADO PROFISSIONAL	6

Fonte: Dados de pesquisa (2024).

A análise do campo AN_INICIO_CURSO identificou registros com valores duplos, como "2002/2010", para programas que incluem mestrado e doutorado. Para corrigir essa questão, foram criados os campos AN_INICIO_CURSO_D e AN_INICIO_CURSO_M, separando os anos de início de cada modalidade. Além disso, foram adicionadas coordenadas geográficas (latitude e longitude) dos municípios onde as IES estão localizadas, permitindo a visualização da distribuição dos programas de pós-graduação e sua proximidade a centros urbanos. Essas modificações padronizam e estruturam os dados, corrigindo limitações e preparando a base para análises futuras.

Para garantir a padronização e a integração dos dados, foram necessários ajustes adicionais, especialmente na atribuição de identificadores numéricos para as áreas do conhecimento, tópico discutido no próximo item.

Ajustes do conjunto de dados dos Programas

Os arquivos da CAPES não possuem identificadores numéricos para as grandes áreas do conhecimento e áreas do conhecimento, apresentando apenas seus nomes. Para inserir os códigos correspondentes, foi desenvolvido um script, conforme apresentado no Quadro 44. Esse *script* realiza uma busca no campo "nome" e insere o código correspondente. Após essa atribuição, o DF4_programas_BR passou a conter 45.587 registros. A inclusão dos códigos permite a integração do DF4_programas_BR com o DF3_trabalhos_BR, estruturando a base de dados para análise.

Quadro 44 – Script desenvolvido para inserção dos códigos nos campos de CD_GRANDE_AREA_CONHECIMENTO e CD_AREA_CONHECIMENTO

```
import pandas as pd
# Dicionários para mapeamento de nome para código
mapa_grande_area = {
'CIÊNCIAS EXATAS E DA TERRA': 10000003, 'CIÊNCIAS BIOLÓGICAS': 20000006, 'ENGENHARIAS': 30000009, 'CIÊNCIAS DA
SAÚDE': 40000001, 'CIÊNCIAS AGRÁRIAS': 50000004, 'CIÊNCIAS SOCIAIS APLICADAS': 60000007, 'CIÊNCIAS HUMANAS':
70000000, 'LINGÜÍSTICA, LETRAS E ARTES': 80000002, 'MULTIDISCIPLINAR': 90000005
}
# O dicionário para mapeamento de nome para código da área
mapa_area = {
'MATEMÁTICA': 10100008, 'PROBABILIDADE E ESTATÍSTICA': 10200002, 'CIÊNCIA DA COMPUTAÇÃO': 10300007,
'ASTRONOMIA': 10400001, 'FÍSICA': 10500006, 'QUÍMICA': 10600000, 'GEOCIÊNCIAS': 10700005, 'OCEANOGRAFIA': 10800000,
'BIOLOGIA GERAL': 20100000, 'GENÉTICA': 20200005, 'BOTÂNICA': 20300000, 'ZOOLOGIA': 20400004, 'ECOLOGIA': 20500009,
'MORFOLOGIA': 20600003, 'FISIOLOGIA': 20700008, 'BIOQUÍMICA': 20800002, 'BIOFÍSICA': 20900007, 'FARMACOLOGIA':
21000000, 'IMUNOLOGIA': 21100004, 'MICROBIOLOGIA': 21200009, 'PARASITOLOGIA': 21300003, 'ENGENHARIA CIVIL':
30100003, 'ENGENHARIA DE MINAS': 30200008, 'ENGENHARIA DE MATERIAIS E METALÚRGICA': 30300002, 'ENGENHARIA
ELÉTRICA': 30400007, 'ENGENHARIA MECÂNICA': 30500001, 'ENGENHARIA QUÍMICA': 30600006, 'ENGENHARIA
SANITÁRIA': 30700000, 'ENGENHARIA DE PRODUÇÃO': 30800005, 'ENGENHARIA NUCLEAR': 30900000, 'ENGENHARIA DE
TRANSPORTES': 31000002, 'ENGENHARIA NAVAL E OCEÂNICA': 31100007, 'ENGENHARIA AEROESPACIAL': 31200001,
'ENGENHARIA BIOMÉDICA': 31300006, 'MEDICINA': 40100006, 'ODONTOLOGIA': 40200000, 'FARMÁCIA': 40300005,
'ENFERMAGEM': 40400000, 'NUTRIÇÃO': 40500004, 'SAÚDE COLETIVA': 40600009, 'FONOAUDIOLOGIA': 40700003,
'FISIOTERAPIA E TERAPIA OCUPACIONAL': 40800008, 'EDUCAÇÃO FÍSICA': 40900002, 'AGRONOMIA': 50100009,
'RECURSOS FLORESTAIS E ENGENHARIA FLORESTAL': 50200003, 'ENGENHARIA AGRÍCOLA': 50300008, 'ZOOTECNIA':
50400002, 'MEDICINA VETERINÁRIA': 50500007, 'RECURSOS PESQUEIROS E ENGENHARIA DE PESCA': 50600001, 'CIÊNCIA
```



```

E TECNOLOGIA DE ALIMENTOS': 50700006, 'DIREITO': 60100001, 'ADMINISTRAÇÃO': 60200006, 'ECONOMIA': 60300000,
'ARQUITETURA E URBANISMO': 60400005, 'PLANEJAMENTO URBANO E REGIONAL': 60500000, 'DEMOGRAFIA': 60600004,
'CIÊNCIA DA INFORMAÇÃO': 60700009, 'MUSEOLOGIA': 60800003, 'COMUNICAÇÃO': 60900008, 'SERVIÇO SOCIAL':
61000000, 'ECONOMIA DOMÉSTICA': 61100005, 'DESENHO INDUSTRIAL': 61200000, 'TURISMO': 61300004, 'FILOSOFIA':
70100004, 'SOCIOLOGIA': 70200009, 'ANTROPOLOGIA': 70300003, 'ARQUEOLOGIA': 70400008, 'HISTÓRIA': 70500002,
'GEOGRAFIA': 70600007, 'PSICOLOGIA': 70700001, 'EDUCAÇÃO': 70800006, 'CIÊNCIA POLÍTICA': 70900000, 'TEOLOGIA':
71000003, 'LINGÜÍSTICA': 80100007, 'LETRAS': 80200001, 'ARTES': 80300006, 'INTERDISCIPLINAR': 90100000, 'ENSINO':
90200000, 'MATERIAIS': 90300009, 'BIOTECNOLOGIA': 90400003, 'CIÊNCIAS AMBIENTAIS': 90500008
}
# Lendo o arquivo DF4.csv
caminho_df4 = "/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF4/DF4.csv"
df4 = pd.read_csv(caminho_df4, sep=";", encoding='UTF-8')
# Mapeando os valores para as novas colunas
df4['CD_GRANDE_AREA_CONHECIMENTO'] = df4['NM_GRANDE_AREA_CONHECIMENTO'].map(mapa_grande_area)
df4['CD_AREA_CONHECIMENTO'] = df4['NM_AREA_CONHECIMENTO'].map(mapa_area)
# Salvando o DataFrame atualizado no arquivo DF4.csv
df4.to_csv(caminho_df4, sep=";", encoding='UTF-8', index=False)

```

Fonte: Dados de pesquisa (2024).

Esses procedimentos expostos no Quadro 44 foram adotados para corrigir limitações dos dados e viabilizar a análise. A adaptação e organização das informações garantem a estrutura necessária para a pesquisa (Pierson, 2021; Mckinney, 2012).

Durante a revisão do DF4_programas_BR, identificou-se a incompletude do campo CD_ENTIDADE_EMEC, resultando em sua exclusão. Já o campo CD_ENTIDADE_CAPES, por estar completo, foi mantido como referência. Para identificar variações nos nomes das instituições associadas a um mesmo código, foi desenvolvido um script, apresentado no Quadro 45.

Quadro 45 – Script desenvolvido para verificação de instituições com nomes diferentes para o mesmo código

```

import pandas as pd
def verificar_consistencia(file_path):
    try:
        # Carrega o DataFrame do arquivo CSV, especificando o delimitador correto
        df = pd.read_csv(file_path, sep=';')
        # Agrupa o DataFrame pelo código da entidade CAPES
        grouped = df.groupby('CD_ENTIDADE_CAPES')
        # Verifica a consistência para cada grupo
        inconsistencias = []
        for name, group in grouped:
            # Verifica se todos os valores em SG_ENTIDADE_ENSINO são iguais dentro do grupo
            if not group['SG_ENTIDADE_ENSINO'].nunique() == 1:
                inconsistencias.append((name, 'SG_ENTIDADE_ENSINO'))
            # Verifica se todos os valores em NM_ENTIDADE_ENSINO são iguais dentro do grupo
            if not group['NM_ENTIDADE_ENSINO'].nunique() == 1:
                inconsistencias.append((name, 'NM_ENTIDADE_ENSINO'))
        # Imprime o relatório
        if inconsistencias:
            print("Foram encontradas inconsistências nos seguintes campos e códigos de entidade CAPES:")
            for codigo, campo in inconsistencias:
                print(f"Código CAPES: {codigo}, Campo: {campo}")
        else:
            print("Todos os campos são consistentes para cada código de entidade CAPES.")
    except Exception as e:
        print(f'Ocorreu um erro ao tentar ler o arquivo: {e}')
file_path = "/content/drive/MyDrive/CAPES/datasets/DF4/DF4.csv"
verificar_consistencia(file_path)

```

Fonte: Dados de pesquisa (2024).

Os resultados da aplicação do script apontaram a necessidade de padronização dos nomes das instituições de ensino, eliminando discrepâncias. Essa padronização, alinhada aos códigos únicos do CD_ENTIDADE_CAPES, assegura a consistência dos dados e a preparação adequada para análises futuras.

Para essa padronização, foi desenvolvido o script do Quadro 47, que lê o DF4_programas_BR e atualiza os campos SG_ENTIDADE_ENSINO e NM_ENTIDADE_ENSINO com base nos valores correspondentes ao CD_ENTIDADE_CAPES. A análise apresentada no Quadro 46 identificou variações nos nomes das instituições associadas ao mesmo código, tornando essa padronização essencial para manter a integridade dos dados.

Quadro 46 – Script desenvolvido para ajustar de instituições com nomes diferentes para o mesmo código

```
import pandas as pd
def atualizar_entidades_ensino(file_path, atualizacoes):
    try:
        # Carrega o DataFrame do arquivo CSV, especificando o delimitador correto
        df = pd.read_csv(file_path, sep=',')
        # Itera sobre a lista de atualizações e aplica cada uma ao DataFrame
        for atualizacao in atualizacoes:
            cd_entidade_capes = atualizacao['CD_ENTIDADE_CAPES']
            sg_entidade_ensino = atualizacao['SG_ENTIDADE_ENSINO']
            nm_entidade_ensino = atualizacao['NM_ENTIDADE_ENSINO']
            df.loc[df['CD_ENTIDADE_CAPES'] == int(cd_entidade_capes), 'SG_ENTIDADE_ENSINO'] = sg_entidade_ensino
            df.loc[df['CD_ENTIDADE_CAPES'] == int(cd_entidade_capes), 'NM_ENTIDADE_ENSINO'] = nm_entidade_ensino
        # Salva o DataFrame modificado de volta ao arquivo CSV
        df.to_csv(file_path, index=False, sep=',')
        print("Os dados foram atualizados e salvos com sucesso.")
    except Exception as e:
        print(f"Ocorreu um erro ao tentar ler ou salvar o arquivo: {e}")
# Exemplo de uso da função com múltiplas atualizações
file_path = "/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF4/DF4.csv"
atualizacoes = [
    {
        'CD_ENTIDADE_CAPES': ,
        'SG_ENTIDADE_ENSINO':,
        'NM_ENTIDADE_ENSINO':},
]
atualizar_entidades_ensino(file_path, atualizacoes)
```

Fonte: Dados de pesquisa (2024).

Aplicando o script do Quadro 46, a padronização dos campos SG_ENTIDADE_ENSINO e NM_ENTIDADE_ENSINO foi concluída, assegurando a consistência dos dados e facilitando a integração e interpretação das informações no DF4_programas_BR.

Dando continuidade ao refinamento do DF4_programas_BR, o Quadro 47 apresenta o script utilizado para a adição dos campos Latitude e Longitude. As coordenadas geográficas

foram coletadas manualmente na plataforma Google Maps¹³, com base nos registros da CAPES, garantindo a incorporação precisa das informações geoespaciais ao conjunto de dados.

Quadro 47 – Script para criar os campos de latitude e longitude ao DF4_programas_BR

```
import pandas as pd
# Caminho para o seu arquivo CSV
file_path = '/content/drive/MyDrive/CAPES/datasets/DF4/DF4.csv'
# Carrega o DataFrame
df = pd.read_csv(file_path, sep=';', encoding='UTF-8')
# Cria colunas LATITUDE e LONGITUDE com valores padrão "NAO DEFINIDO"
df['LATITUDE'] = "NAO DEFINIDO"
df['LONGITUDE'] = "NAO DEFINIDO"
# Salva o DataFrame com as novas colunas
df.to_csv(file_path, index=False, sep=';', encoding='UTF-8')
print("Colunas LATITUDE e LONGITUDE criadas e arquivo salvo.")
```

Fonte: Dados de pesquisa (2024).

Algumas IES possuem campi avançados, estendendo sua atuação para além da sede principal. No entanto, nos conjuntos de dados da CAPES, a identificação ocorre apenas pelo código CD_ENTIDADE_CAPES. Para suprir essa limitação, foram coletadas as coordenadas geográficas dos municípios listados na base de dados da CAPES. Para as IES que não especificam a distribuição de cursos entre diferentes localidades, foram adotadas as coordenadas do município da sede principal. O Quadro 48 apresenta o procedimento para a inserção desses valores.

Quadro 48 – Script para inserção de valores de latitude e longitude aos campos do DF4_programas_BR

```
import pandas as pd
# Caminho para o seu arquivo CSV
file_path = '/content/drive/MyDrive/CAPES/datasets/DF4/DF4.csv'
# Carrega o DataFrame
df = pd.read_csv(file_path, sep=';', encoding='UTF-8')
# Dicionário com algumas coordenadas de exemplo (latitude, longitude)
coordenadas = {
    'EXEMPLO_01': (-1.72183, -48.8788),
    'EXEMPLO_02': (-4.22083, -38.7055),
    'EXEMPLO_03': (-12.1356, -38.4192),
    # Adicione mais municípios aqui
}
# Atualiza as colunas LATITUDE e LONGITUDE com as coordenadas
for municipio, (lat, lon) in coordenadas.items():
    df.loc[df['NM_MUNICIPIO_PROGRAMA_IES'] == municipio, 'LATITUDE'] = lat
    df.loc[df['NM_MUNICIPIO_PROGRAMA_IES'] == municipio, 'LONGITUDE'] = lon
# Salva o DataFrame modificado
df.to_csv(file_path, index=False, sep=';', encoding='UTF-8')
print("Dados de LATITUDE e LONGITUDE inseridos e arquivo salvo.")
```

Fonte: Dados de pesquisa (2024).

¹³ <https://www.google.com.br/maps>

Na sequência da análise, o arquivo CSV do DF4_programas_BR foi aberto no LibreOffice Calc para inspeção visual. O procedimento revelou que algumas IES são identificadas por um único código de entidade (CD_ENTIDADE_CAPES), embora estejam associadas a múltiplos municípios em seus programas. Diante disso, foi decidido ajustar os campos de nome e sigla da entidade, seguindo a prática já adotada pela CAPES para algumas instituições. O código de entidade CAPES foi mantido inalterado para garantir sua integração com outras bases de dados. O Quadro 49 apresenta o script desenvolvido para listar as IES associadas a mais de um município com o mesmo código.

Quadro 49 – Script para retornar as IES com mais de um município o mesmo

```
import pandas as pd
# Caminho para o seu arquivo CSV
file_path = '/content/drive/MyDrive/CAPES/datasets/DF4/DF4.csv'
# Carrega o DataFrame
df = pd.read_csv(file_path, sep=';', encoding='UTF-8')
# Agrupa o DataFrame pelos campos CD_ENTIDADE_CAPES e NM_ENTIDADE_ENSINO
grouped = df.groupby(['CD_ENTIDADE_CAPES', 'NM_ENTIDADE_ENSINO'])
# Inicializa contadores
casos_multiplos_municipios = 0
casos_unicos = 0
# Lista para armazenar os casos com múltiplos municípios
entidades_multiplos_municipios = []
for (cd_entidade, nm_entidade), group in grouped:
    municipios_unicos = group['NM_MUNICIPIO_PROGRAMA_IES'].unique()
    if len(municipios_unicos) > 1:
        # Caso a entidade tenha mais de um município associado
        casos_multiplos_municipios += 1
        entidades_multiplos_municipios.append((cd_entidade, nm_entidade, len(municipios_unicos)))
    else:
        # Caso todos os campos sejam iguais (considerando apenas um município)
        casos_unicos += 1
# Exibe os resultados
print(f"Casos com mais de um município: {casos_multiplos_municipios}")
for entidade in entidades_multiplos_municipios:
    print(f"Código: {entidade[0]}, Nome: {entidade[1]}, Municípios: {entidade[2]}")
print(f"Casos com todos os campos iguais: {casos_unicos}")
```

Fonte: Dados de pesquisa (2024).

A análise identificou 97 casos em que um mesmo código e nome de IES estão associados a mais de um município, enquanto em 405 casos essa correspondência não ocorre. Um exemplo é a UNIVERSIDADE DO ESTADO DA BAHIA, vinculada a dez municípios. Para IES com múltiplos códigos de entidade na CAPES, observa-se um padrão na nomenclatura e sigla, como em UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO (NOVA IGUAÇU) e UFRRJ-NI. Nesse modelo, o nome da IES é seguido pelo município entre parênteses, enquanto a sigla pode variar, adotando, em alguns casos, as iniciais do município para diferenciação. Com base nessa análise, foi desenvolvido um *script*, apresentado no Quadro 50.

Quadro 50 – Script desenvolvido para inserir o nome do município ao nome da IES

```
import pandas as pd
# Define o caminho para o arquivo CSV
file_path = '/content/drive/MyDrive/CAPES/datasets/DF4/DF4.csv'
# Carrega o DataFrame do arquivo CSV
df = pd.read_csv(file_path, sep=';', encoding='UTF-8')
# Agrupa o DataFrame por CD_ENTIDADE_CAPES e NM_ENTIDADE_ENSINO
grouped = df.groupby(['CD_ENTIDADE_CAPES', 'NM_ENTIDADE_ENSINO'])
# Prepara um DataFrame para as alterações
df_alterado = df.copy()
# Contador para as alterações realizadas
alteracoes = 0
for (cd_entidade, nm_entidade), group in grouped:
    municipios_unicos = group['NM_MUNICIPIO_PROGRAMA_IES'].unique()
    if len(municipios_unicos) > 1:
        for municipio in municipios_unicos:
            # Identifica os índices dos registros a serem alterados
            indices = group[group['NM_MUNICIPIO_PROGRAMA_IES'] == municipio].index
            # Atualiza NM_ENTIDADE_ENSINO para incluir o nome do município
            for indice in indices:
                nm_entidade_atualizado = f'{nm_entidade} ( {municipio} )'
                df_alterado.at[indice, 'NM_ENTIDADE_ENSINO'] = nm_entidade_atualizado
                alteracoes += 1
            print(f'Alterado: {nm_entidade} para {nm_entidade_atualizado}')
# Verifica se houve alterações e salva o arquivo
if alteracoes > 0:
    df_alterado.to_csv(file_path, index=False, sep=';', encoding='UTF-8')
    print(f'\nTotal de alterações realizadas: {alteracoes}')
else:
    print("Nenhuma alteração necessária foi identificada.")
```

Fonte: Dados de pesquisa (2024).

O *script* do Quadro 50 foi utilizado para identificar casos em que um mesmo CD_ENTIDADE_CAPES está associado a mais de um NM_MUNICIPIO_PROGRAMA_IES. O script agrupa o DF4_programas_BR por CD_ENTIDADE_CAPES e NM_ENTIDADE_ENSINO, verificando se há múltiplos municípios associados. Quando isso ocorre, o NM_ENTIDADE_ENSINO é atualizado para incluir o nome do município correspondente. Essa padronização melhora a representação das IES nos dados e possibilita uma análise mais precisa da distribuição geográfica dos programas de pós-graduação.

Para o campo SG_ENTIDADE_ENSINO (sigla da IES), um *script* foi inicialmente elaborado para automatizar as alterações. No entanto, testes indicaram resultados insatisfatórios. Como alternativa, optou-se pela edição manual no LibreOffice Calc, permitindo inspeção e correção direta das siglas associadas às IES. Esse processo garantiu a correta identificação dos programas de pós-graduação por município, preservando a integridade dos dados.

No refinamento do DF4_programas_BR, o foco seguinte foi o campo AN_INICIO_CURSO, que registra os anos de início dos programas de mestrado e doutorado em uma única entrada. Para separá-los, foram criados os campos AN_INICIO_CURSO_M e

AN_INICIO_CURSO_D, correspondendo aos anos de início do mestrado e do doutorado, respectivamente.

Um *script* foi desenvolvido para verificar a existência desses novos campos no DataFrame e criá-los com valores nulos, caso não estivessem presentes. Em seguida, foi implementada uma função que analisa cada entrada do campo AN_INICIO_CURSO, identificando registros com um único ano ou dois anos separados por "/". A função distribui os valores corretamente entre os novos campos. A metodologia e o script correspondente são apresentados no Quadro 51.

Quadro 51 – Script para separar os anos de início dos cursos de mestrado e doutorado

```
import pandas as pd
# Define o caminho para o arquivo CSV
file_path = '/content/drive/MyDrive/CAPEs/datasets/DF4/DF4.csv'
# Carrega o DataFrame do arquivo CSV
df = pd.read_csv(file_path, sep=';', encoding='UTF-8')
# Verifica se os campos necessários existem ou os cria
if 'AN_INICIO_CURSO_D' not in df.columns:
    df['AN_INICIO_CURSO_D'] = None
if 'AN_INICIO_CURSO_M' not in df.columns:
    df['AN_INICIO_CURSO_M'] = None
# Função para processar cada valor do campo AN_INICIO_CURSO
def processar_an_inicio_curso(valor):
    if '/' in valor:
        anos = valor.split('/')
        return min(anos), max(anos)
    else:
        return valor, None
# Aplica a função ao DataFrame
df[['AN_INICIO_CURSO_M', 'AN_INICIO_CURSO_D']] = df['AN_INICIO_CURSO'].apply(
    lambda x: pd.Series(processar_an_inicio_curso(x))
)
# Salva o DataFrame modificado de volta ao arquivo CSV
df.to_csv(file_path, index=False, sep=';', encoding='UTF-8')
print("Processamento concluído e arquivo salvo.")
```

Fonte: Dados de pesquisa (2024).

A execução do *script* organiza a estrutura dos dados, permitindo análises específicas para cada nível de formação e aplicando técnicas de manipulação para estruturar grandes conjuntos de dados acadêmicos.

Na continuidade da estruturação do DF4_programas_BR, o foco agora é a padronização do campo NM_GRAU_PROGRAMA. A análise dos dados indicou a possibilidade de atribuir um código numérico a cada categoria desse campo, melhorando a eficiência das consultas. Essa conversão substitui as designações textuais dos graus acadêmicos por valores codificados, conforme apresentado na Tabela 17. A implementação dessa transformação foi realizada por meio de um *script*, detalhado no Quadro 52.

Quadro 52 – Script para criar códigos para o nome do grau do programa.

```
import pandas as pd
# Define o caminho para o arquivo CSV
file_path = '/content/drive/MyDrive/CAPES/datasets/DF4/DF4.csv'
# Carrega o DataFrame do arquivo CSV
df = pd.read_csv(file_path, sep=';', encoding='UTF-8')
# Mapeamento dos valores de NM_GRAU_PROGRAMA para CD_GRAU_PROGRAMA
mapeamento_grau = {'MESTRADO': 1, 'MESTRADO/DOUTORADO': 2, 'DOUTORADO': 3, 'MESTRADO PROFISSIONAL': 4, 'DOUTORADO PROFISSIONAL': 5, 'MESTRADO PROFISSIONAL/DOUTORADO PROFISSIONAL': 6,}
# Cria o campo CD_GRAU_PROGRAMA com valores nulos inicialmente
df.insert(df.columns.get_loc('NM_PROGRAMA_IDIOMA') + 1, 'CD_GRAU_PROGRAMA', None)
# Preenche o campo CD_GRAU_PROGRAMA com os valores correspondentes
df['CD_GRAU_PROGRAMA'] = df['NM_GRAU_PROGRAMA'].map(mapeamento_grau)
# Salva o DataFrame modificado de volta ao arquivo CSV
df.to_csv(file_path, index=False, sep=';', encoding='UTF-8')
print("Campo CD_GRAU_PROGRAMA criado e preenchido com sucesso.")
```

Fonte: Dados de pesquisa (2024).

Após a implementação do *script* no Quadro 52, o campo CD_GRAU_PROGRAMA foi adicionado e preenchido corretamente, e essa transformação padroniza os graus acadêmicos, facilitando as consultas e análises comparativas entre os diferentes níveis de formação.

Dando continuidade à pesquisa, a análise foi direcionada para os campos CD_AREA_AVALIACAO e NM_AREA_AVALIACAO, que representam as áreas de avaliação das IES. Esperava-se uma correspondência direta entre os valores desses campos, mas a análise inicial identificou oito nomes adicionais sem códigos equivalentes. Para quantificar essa discrepância, foi desenvolvido um script, apresentado no Quadro 53.

Quadro 53 – Script para verificação dos códigos e nomes das áreas de avaliação

```
import pandas as pd
# Define o caminho para o arquivo CSV
file_path = '/content/drive/MyDrive/CAPES/datasets/DF4/DF4.csv'
# Carrega o DataFrame do arquivo CSV
df = pd.read_csv(file_path, sep=';', encoding='UTF-8')
# Obtém listas únicas de códigos e nomes de áreas de avaliação
unique_cd_area = df['CD_AREA_AVALIACAO'].unique()
unique_nm_area = df['NM_AREA_AVALIACAO'].unique()
# Converte códigos para string, caso não sejam
unique_cd_area = set([str(cd) for cd in unique_cd_area])
# Cria um dicionário para mapear códigos para nomes
codigo_para_nome = df.drop_duplicates(subset=['CD_AREA_AVALIACAO'])\
    .set_index('CD_AREA_AVALIACAO')['NM_AREA_AVALIACAO'].to_dict()
# Identifica nomes sem códigos correspondentes
nomes_sem_codigo = [nome for nome in unique_nm_area if nome not in codigo_para_nome.values()]
# Exibe os resultados
if nomes_sem_codigo:
    print("Nomes de áreas de avaliação sem códigos correspondentes:")
    for nome in nomes_sem_codigo:
        print(nome)
else:
    print("Todos os nomes de áreas de avaliação possuem códigos correspondentes.")
```

Fonte: Dados de pesquisa (2024).

A execução do *script* confirmou a existência desses oito nomes sem correspondência com códigos, validando a inconsistência observada. Em seguida, foi realizada uma

comparação com as informações oficiais da CAPES (Capes, 2024), utilizado como referência para as áreas de avaliação.

Os nomes discrepantes estão listados na Tabela 18, uma vez que para facilitar a análise, os dados foram convertidos para o formato de planilha, permitindo a identificação das inconsistências e orientando o processo de correção.

Tabela 18 – Nomes das áreas de avaliação que são diferentes dos códigos

Código	Nome correto	Nome errado
27	ADMINISTRAÇÃO PÚBLICA E DE EMPRESAS, CIÊNCIAS CONTÁBEIS E TURISMO	ADMINISTRAÇÃO, CIÊNCIAS CONTÁBEIS E TURISMO
29	ARQUITETURA, URBANISMO E DESIGN	ARQUITETURA E URBANISMO
11	ARTES	ARTES / MÚSICA
31	COMUNICAÇÃO E INFORMAÇÃO	CIÊNCIAS SOCIAIS APLICADAS I
44	CIÊNCIAS DA RELIGIÃO E TEOLOGIA	FILOSOFIA/TEOLOGIA:SUBCOMISSÃO TEOLOGIA
44	CIÊNCIAS DA RELIGIÃO E TEOLOGIA	TEOLOGIA
33	FILOSOFIA	FILOSOFIA / TEOLOGIA:SUBCOMISSÃO FILOSOFIA
41	LINGUÍSTICA E LITERATURA	LETRAS / LINGUÍSTICA

Fonte: Dados de pesquisa (2024).

O processo identificou áreas de avaliação que exigiam ajustes nos registros e destacou a importância da verificação cruzada com fontes oficiais para garantir a precisão dos dados. A validação dessas informações é essencial em pesquisas que utilizam dados institucionais complexos (Nanda *et al.*, 2000).

Após a revisão do DF4_programas_BR, a análise focou nos campos CD_PROGRAMA_IES e NM_PROGRAMA_IES, que representam o código e o nome dos programas de pós-graduação, conforme registrados pela IES e validados pelo Ministério da Educação. Identificou-se uma diferença de 2.662 registros entre esses campos, sugerindo que deveriam possuir a mesma quantidade de entradas. A inspeção visual no LibreOffice Calc revelou que alguns programas possuíam múltiplos códigos, mas mantinham o mesmo nome. Essa variação indica inconsistências na correspondência entre código e nome do programa, impossibilitando uma padronização direta. Para diferenciá-los, pode-se adotar o código do programa como identificador único ou concatenar a sigla da IES ao nome do programa. Essa abordagem permite distinguir programas de pós-graduação com maior precisão e reflete a diversidade das ofertas aprovadas pelo Ministério da Educação.

A otimização do DF4_programas_BR, baseada nos dados da CAPES, evidencia a necessidade de precisão e integridade na pesquisa acadêmica. A manipulação de grandes

volumes de dados exige estratégias adequadas de análise e organização. Como aponta Kitchin (2014), a "data revolution" não se restringe ao aumento da disponibilidade de dados, mas também envolve o aprimoramento das técnicas de análise e gestão para garantir sua validade e aplicabilidade.

A validação cruzada com fontes oficiais reforça a importância da confiabilidade dos dados em investigações científicas. Boyd e Crawford (2012) ressaltam que a integridade dos dados não é apenas uma questão técnica, mas também um compromisso ético na produção do conhecimento.

Com essas correções, o DF4_programas_BR agora pode contribuir para o aprimoramento do DF3_trabalhos_BR. O refinamento desse conjunto de dados possibilita uma análise mais precisa dos programas de pós-graduação no Brasil e fortalece a discussão sobre práticas de gestão de dados na pesquisa acadêmica. A documentação deste processo dialoga com debates sobre ciência de dados, ressaltando a importância de abordagens rigorosas e éticas no uso de grandes volumes de informações digitais.

Após a consolidação dos dados dos programas de pós-graduação, a pesquisa avança para a análise dos docentes vinculados aos Programas de Pós-Graduação (PPG) *stricto sensu* no Brasil. A CAPES disponibiliza conjuntos de dados referentes ao período de 2004 a 2022, apresentando variações na estrutura e no conteúdo ao longo dos anos.

Passo 2: Entendimento e ajustes do conjunto de dados dos Docentes

De 2004 a 2012, as informações estão concentradas em um único arquivo, contendo 491.773 registros com dados sobre nome, categoria, maior titulação e vínculo institucional dos docentes. Já de 2013 a 2021, os dados são distribuídos em nove arquivos anuais, totalizando 881.639 registros, com detalhes adicionais sobre nível de titulação acadêmica e atuação docente. Diante dessas diferenças, é criado um novo DataFrame DF5_docentes_BR, que reunirá as informações mais relevantes de ambos os períodos.

O DF5_docentes_BR incluirá os seguintes campos: Ano-base, Código e nome do programa, Identificação do docente, País de origem, Ano de nascimento, Categoria do docente (permanente ou colaborador), Natureza do vínculo (servidor público, colaborador, bolsista, etc.), Titulação atual e ano da titulação, Sigla e nome da entidade, País da entidade, Área do conhecimento e código da área do conhecimento.

Na etapa inicial do tratamento dos dados, cada arquivo é analisado para compreender sua estrutura e garantir a correta integração das informações no DF5_docentes_BR. Esse

processo permite consolidar os dados de forma organizada, assegurando que as análises futuras sejam consistentes e representativas. A seguir apresentam-se o entendimento (item 1) e os ajustes (item 2) do conjunto de dados dos docentes.

Entendimento do conjunto de dados dos Docentes

A análise dos dados dos docentes começa com a comparação entre os conjuntos referentes aos períodos de 2004 a 2012 e 2013 a 2022. O primeiro conjunto contém 33 colunas, enquanto o segundo apresenta 41 colunas, indicando a adição de oito novas variáveis ao longo do tempo. Essa diferença exige a padronização dos dados para assegurar a integridade e a comparabilidade dos registros.

A Tabela 19 exibe o volume total por campo em ambos os conjuntos de dados, permitindo identificar variáveis comuns e divergentes ao longo dos períodos analisados. Essa análise orienta a estruturação do DF5_docentes_BR, garantindo a correta consolidação das informações dos docentes.

Tabela 19 – Total dos dados dos conjuntos de dados referentes aos docentes

2004-2012		2013-2022	
Campo	Linhas	Campo	Linhas
AN_BASE	491773	AN_BASE	991187
SG_ENTIDADE_ENSINO	491773	CD_AREA_AVALIACAO	991187
NM_ENTIDADE_ENSINO	491773	NM_AREA_AVALIACAO	991187
CS_STATUS_JURIDICO	491773	NM_GRANDE_AREA_CONHECIMENTO	991187
NM_REGIAO_ENTIDADE	491773	NM_AREA_CONHECIMENTO	991187
SG_UF_ENTIDADE_ENSINO	491773	CD_PROGRAMA_IES	991187
CD_PROGRAMA_IES	491773	NM_PROGRAMA_IES	991187
NM_PROGRAMA_IES	491773	NM_GRAU_PROGRAMA	991187
NM_MODALIDADE_PROGRAMA	491773	NM_MODALIDADE_PROGRAMA	991187
NM_NIVEL_PROGRAMA	491773	CD_CONCEITO_PROGRAMA	991187
ID_AREA_AVALIACAO	491773	CD_ENTIDADE_CAPES	991187
NM_AREA_AVALIACAO	491773	CD_ENTIDADE_EMEC	989010
NR_SEQUENCIAL_DOCENTE	491773	SG_ENTIDADE_ENSINO	991187
NM_DOCENTE	491773	NM_ENTIDADE_ENSINO	991187
NM_PAIS_ORIGEM_DOCENTE	491773	DS_DEPENDENCIA_ADMINISTRATIVA	991187
AN_NASCIMENTO_DOCENTE	491773	CS_STATUS_JURIDICO	991187

2004-2012		2013-2022	
Campo	Linhas	Campo	Linhas
NR_IDADE	491512	NM_MUNICIPIO_PROGRAMA_IES	991187
DS_FAIXA_ETARIA	491512	SG_UF_PROGRAMA	991187
DS_CATEGORIA_DOCENTE	491773	NM_REGIAO	991187
DS_TIPO_NATUR_VINCULO_DOCENTE	491773	ID_PESSOA	991187
AN_INICIO_VINCULO_DOCENTE	491773	TP_DOCUMENTO_DOCENTE	991187
ME_INICIO_VINCULO_DOCENTE	491773	NR_DOCUMENTO_DOCENTE	991187
ME_FIM_VINCULO_DOCENTE	491773	NM_DOCENTE	991187
IN_AFASTAMENTO_DOCENTE	26012	AN_NASCIMENTO_DOCENTE	991179
DS_MOTIVO_AFASTAMENTO_DOCENTE	491773	DS_FAIXA_ETARIA	991084
DS_TITULACAO_ATUAL_DOCENTE	491773	DS_TIPO_NACIONALIDADE_DOCENTE	991187
AN_TITULACAO_DOCENTE	491773	NM_PAIS_NACIONALIDADE_DOCENTE	991187
SG_IES_TIT_MAX_DOCENTE	491773	DS_CATEGORIA_DOCENTE	991187
NM_IES_TIT_MAX_DOCENTE	491773	DS_TIPO_VINCULO_DOCENTE_IES	991168
NM_PAIS_IES_TIT_MAX_DOCENTE	491773	DS_REGIME_TRABALHO	991187
NM_AREA_CONHEC_FORM_DOCENTE	491773	CD_CAT_BOLSA_PRODUTIVIDADE	227245
SG_IES_ORIGEM_DOCENTE	491773	IN_DOUTOR	991187
NM_IES_ORIGEM_DOCENTE	491773	AN_TITULACAO	991173
		NM_GRAU_TITULACAO	991173
		CD_AREA_BASICA_TITULACAO	991173
		NM_AREA_BASICA_TITULACAO	991173
		SG_IES_TITULACAO	977335
		NM_IES_TITULACAO	991173
		NM_PAIS_IES_TITULACAO	991173
		ID_ADD_FOTO_PROGRAMA	991187
		ID_ADD_FOTO_PROGRAMA_IES	991187

Fonte: Dados de pesquisa (2024).

A análise dos conjuntos de dados da CAPES identificou variáveis como área de avaliação, área do conhecimento, status jurídico da instituição, grau do programa, modalidade do programa, conceito do programa, município, UF e região. Embora essas informações

sejam relevantes, estão detalhadas em outros DataFrames e, para evitar redundância, foram excluídas do DF5_docentes_BR, que será focado exclusivamente nos docentes dos programas de pós-graduação *stricto sensu*.

Além disso, algumas variáveis presentes apenas no conjunto de dados mais recente, como o identificador da foto do programa, não possuem correspondência nos períodos anteriores. Para garantir a consistência temporal e a comparabilidade dos dados, essas variáveis não foram incorporadas ao DF5_docentes_BR.

A Tabela 20 apresenta a composição final do DF5_docentes_BR, detalhando os campos selecionados e o critério adotado na filtragem das informações, assegurando que a análise reflita com precisão o perfil dos docentes nos programas de pós-graduação avaliados pela CAPES.

Tabela 20 – Campos que vão constituir o DF5_docentes_BR

Campo
ANO_BASE
CD_PROGRAMA
NM_PROGRAMA
ID_DOCENTE
NM_DOCENTE
NM_NACIONALIDADE_DOCENTE
AN_NASCIMENTO_DOCENTE
DS_CATEGORIA_DOCENTE
DS_TIPO_VINCULO_DOCENTE
NM_ULTIMO_GRAU_TITULACAO
AN_ULTIMO_GRAU_TITULACAO
SG_IES_ULTIMO_GRAU_TITULACAO
NM_IES_GRAU_TITULACAO
NM_PAIS_ULTIMO_GRAU_TITULACAO
CD_AREA_BASICA_TITULACAO
NM_AREA_BASICA_TITULACAO

Fonte: Dados de pesquisa (2024).

Prosseguindo a análise, é construído o DataFrame DF5_docentes_BR, cujos procedimentos estão no script do Quadro 54.

Quadro 54 – Script desenvolvido para alteração do encoding e separador dos arquivos

```
import pandas as pd
# Criar um DataFrame vazio com as colunas especificadas
colunas = ["ANO_BASE", "CD_PROGRAMA", "NM_PROGRAMA", "ID_DOCENTE", "NM_DOCENTE",
"NM_NACIONALIDADE_DOCENTE", "AN_NASCIMENTO_DOCENTE", "DS_CATEGORIA_DOCENTE",
"DS_TIPO_VINCULO_DOCENTE", "NM_ULTIMO_GRAU_TITULACAO", "AN_ULTIMO_GRAU_TITULACAO",
"SG_IES_ULTIMO_GRAU_TITULACAO", "NM_IES_GRAU_TITULACAO", "NM_PAIS_ULTIMO_GRAU_TITULACAO",
"CD_AREA_BASICA_TITULACAO", "NM_AREA_BASICA_TITULACAO"
]
df5 = pd.DataFrame(columns=colunas)
# Salvar o DataFrame como um arquivo CSV
df5.to_csv("DF5.csv", index=False, encoding="utf-8", sep=";")
```

Fonte: Dados de pesquisa (2024).

Após a execução do script, os dados são incorporados ao DF5_docentes_BR, seguido de uma revisão para ajustes necessários. Esse processo organiza as informações e possibilita análises sobre o corpo docente nos programas de pós-graduação avaliados.

A estrutura do DF5_docentes_BR foi definida com campos específicos para registrar os dados dos docentes. A incorporação das informações ocorreu em duas etapas, sendo a primeira referente ao período de 2004 a 2012. O Quadro 55 apresenta o script utilizado nessa inserção, que incluiu a conversão do *encoding* dos arquivos da CAPES, originalmente em ISO-8859-1, para UTF-8, garantindo compatibilidade e integridade na manipulação dos dados. Esse procedimento assegura a correta estruturação do DF5_docentes_BR, preparando-o para as análises acadêmicas subsequentes.

Quadro 55 – Script desenvolvido para alteração do encoding e separador dos arquivos

```

import pandas as pd
# Caminho do arquivo de entrada
input_path = "/content/drive/MyDrive/CAPES/datasets/DF5/br-capes-colsucup-docente-2004-2012.csv"
# Caminho do arquivo DF5.csv
df5_path = "/content/drive/MyDrive/CAPES/datasets/DF5/DF5.csv"
# Ler o arquivo CSV de entrada
df_input = pd.read_csv(input_path, encoding="ISO-8859-1", sep=";")
# Mapeamento de->para
mapping={"ANO_BASE":"ANO_BASE","CD_PROGRAMA_IES":"CD_PROGRAMA","NM_PROGRAMA_IES":
"NM_PROGRAMA","NM_DOCENTE":"NM_DOCENTE","NM_PAIS_ORIGEM_DOCENTE":
"NM_NACIONALIDADE_DOCENTE","AN_NASCIMENTO_DOCENTE":"AN_NASCIMENTO_DOCENTE",
"DS_CATEGORIA_DOCENTE":"DS_CATEGORIA_DOCENTE","DS_TIPO_NATUR_VINCULO_DOCENTE":
"DS_TIPO_VINCULO_DOCENTE","DS_TITULACAO_ATUAL_DOCENTE":"NM_ULTIMO_GRAU_TITULACAO",
"AN_TITULACAO_DOCENTE":"AN_ULTIMO_GRAU_TITULACAO","SG_IES_TIT_MAX_DOCENTE":
"SG_IES_ULTIMO_GRAU_TITULACAO","NM_IES_TIT_MAX_DOCENTE":"NM_IES_GRAU_TITULACAO",
"NM_PAIS_IES_TIT_MAX_DOCENTE":"NM_PAIS_ULTIMO_GRAU_TITULACAO",
"NM_AREA_CONHEC_FORM_DOCENTE":"NM_AREA_BASICA_TITULACAO"}
# Adicionar colunas com valor 'null' para ID_DOCENTE e CD_AREA_BASICA_TITULACAO
df_input["ID_DOCENTE"] = 'null'
df_input["CD_AREA_BASICA_TITULACAO"] = 'null'
# Renomear as colunas conforme o mapeamento
df_input.rename(columns=mapping, inplace=True)
# Reordenar as colunas conforme especificado
columns_order = ["ANO_BASE", "CD_PROGRAMA", "NM_PROGRAMA", "ID_DOCENTE", "NM_DOCENTE",
"NM_NACIONALIDADE_DOCENTE", "AN_NASCIMENTO_DOCENTE",
"DS_CATEGORIA_DOCENTE", "DS_TIPO_VINCULO_DOCENTE", "NM_ULTIMO_GRAU_TITULACAO",
"AN_ULTIMO_GRAU_TITULACAO", "SG_IES_ULTIMO_GRAU_TITULACAO", "NM_IES_GRAU_TITULACAO",
"NM_PAIS_ULTIMO_GRAU_TITULACAO", "CD_AREA_BASICA_TITULACAO", "NM_AREA_BASICA_TITULACAO"]
df_mapped = df_input[columns_order]
# Ler o arquivo DF5.csv
df5 = pd.read_csv(df5_path, sep=";")
# Adicionar os dados mapeados ao DF5.csv
df5 = pd.concat([df5, df_mapped], ignore_index=True)
# Salvar o DataFrame atualizado no arquivo DF5.csv
df5.to_csv(df5_path, index=False, encoding="utf-8", sep=";")
print("Dados inseridos com sucesso em", df5_path)

```

Fonte: Dados de pesquisa (2024).

O *script* do Quadro 55 exigiu um ajuste específico devido à ausência do campo ID_DOCENTE nos registros de 2004 a 2012, período em que apenas o nome do docente está disponível. Para resolver essa limitação, foi gerado um identificador único para cada docente não presente nos dados de 2013 a 2022. A mesma estratégia é aplicada ao código da área básica de titulação, garantindo a integração e continuidade dos registros ao longo do tempo.

Assim, os dados são incorporados ao DF5_docentes_BR, consolidando as informações dos docentes dos programas de pós-graduação. A Tabela 21 apresenta a composição atualizada do DF5_docentes_BR, destacando o processo de padronização e estruturação do conjunto de dados.

Tabela 21 – Dados inseridos no DF5_docentes_BR após a execução do script

Índice	Campo	Linhas	Valores únicos
1	ANO_BASE	491773	9
2	CD_PROGRAMA	491773	3563

Índice	Campo	Linhas	Valores únicos
3	NM_PROGRAMA	491773	1652
4	ID_DOCENTE	0	0
5	NM_DOCENTE	491773	76999
6	NM_NACIONALIDADE_DOCENTE	491773	127
7	AN_NASCIMENTO_DOCENTE	491773	83
8	DS_CATEGORIA_DOCENTE	491773	3
9	DS_TIPO_VINCULO_DOCENTE	491773	4
10	NM_ULTIMO_GRAU_TITULACAO	491773	4
11	AN_ULTIMO_GRAU_TITULACAO	491773	71
12	SG_IES_ULTIMO_GRAU_TITULACAO	491773	6225
13	NM_IES_GRAU_TITULACAO	491773	12990
14	NM_PAIS_ULTIMO_GRAU_TITULACAO	491773	97
15	CD_AREA_BASICA_TITULACAO	0	0
16	NM_AREA_BASICA_TITULACAO	491773	1075

Fonte: Dados de pesquisa (2024).

A Tabela 21 confirma as observações da Tabela 19, servindo como referência para a continuidade da integração dos dados no DF5_docentes_BR. A próxima ação envolve a incorporação dos registros do período de 2013 a 2022, que, apesar de estarem distribuídos em arquivos distintos, seguem a mesma estrutura dos dados anteriores. Para garantir a padronização, é necessária uma conversão na codificação dos arquivos. O Quadro 56 detalha o procedimento para a inserção desses dados.

Quadro 56 – Script desenvolvido para inserção dos dados dos arquivos de 2013 a 2022 no DF5_docentes_BR

```
import pandas as pd
import os
# Caminho da pasta com os arquivos CSV
folder_path = "/content/drive/MyDrive/CAPES/datasets/DF5/"
```

```

# Caminho do arquivo DF5.csv
df5_path = "/content/drive/MyDrive/CAPES/datasets/DF5/DF5.csv"
# Mapeamento de->para
mapping = {
    "AN_BASE": "ANO_BASE",
    "CD_PROGRAMA_IES": "CD_PROGRAMA", "NM_PROGRAMA_IES": "NM_PROGRAMA", "ID_PESSOA": "ID_DOCENTE",
    "NM_DOCENTE": "NM_DOCENTE", "NM_PAIS_NACIONALIDADE_DOCENTE": "NM_NACIONALIDADE_DOCENTE",
    "AN_NASCIMENTO_DOCENTE": "AN_NASCIMENTO_DOCENTE", "DS_CATEGORIA_DOCENTE":
    "DS_CATEGORIA_DOCENTE", "DS_TIPO_VINCULO_DOCENTE_IES": "DS_TIPO_VINCULO_DOCENTE",
    "NM_GRAU_TITULACAO": "NM_ULTIMO_GRAU_TITULACAO", "AN_TITULACAO": "AN_ULTIMO_GRAU_TITULACAO",
    "SG_IES_TITULACAO": "SG_IES_ULTIMO_GRAU_TITULACAO", "NM_IES_TITULACAO": "NM_IES_GRAU_TITULACAO",
    "NM_PAIS_IES_TITULACAO": "NM_PAIS_ULTIMO_GRAU_TITULACAO", "CD_AREA_BASICA_TITULACAO":
    "CD_AREA_BASICA_TITULACAO", "NM_AREA_BASICA_TITULACAO": "NM_AREA_BASICA_TITULACAO" }
# Ler o arquivo DF5.csv existente
df5 = pd.read_csv(df5_path, sep=";", encoding="utf-8")
# Iterar pelos arquivos CSV dos anos de 2013 a 2022
for year in range(2013, 2023):
    file_path = os.path.join(folder_path, f'{year}.csv')
    if os.path.exists(file_path):
        try:
            # Tentar ler o arquivo CSV com codificação utf-8
            df = pd.read_csv(file_path, sep=";", encoding="utf-8")
        except UnicodeDecodeError:
            # Se ocorrer um erro, tentar ler com codificação ISO-8859-1
            df = pd.read_csv(file_path, sep=";", encoding="ISO-8859-1")
        # Renomear as colunas conforme o mapeamento
        df.rename(columns=mapping, inplace=True)
        # Selecionar apenas as colunas de interesse
        df = df[list(mapping.values())]
        # Adicionar os dados mapeados ao DF5.csv
        df5 = pd.concat([df5, df], ignore_index=True)
# Salvar o DataFrame atualizado no arquivo DF5.csv
df5.to_csv(df5_path, index=False, sep=";", encoding="utf-8")

```

Fonte: Dados de pesquisa (2024).

Com a consolidação dos dados no DF5_docentes_BR, inicia-se a análise do conjunto unificado. Isso permite identificar possíveis inconsistências ou ajustes que não eram perceptíveis nos conjuntos fragmentados. A Tabela 22 apresenta uma visão geral do DF5_docentes_BR, possibilitando a avaliação das informações organizadas. Os registros dos docentes, abrangendo o período de 2004 a 2022, confirmam a existência de 19 valores distintos para os anos analisados e um total de 1.482.960 linhas. Esses números servem como referência para a análise dos demais campos e a validação da estrutura consolidada do conjunto de dados.

Tabela 22 – Dados inseridos no DF5 após a execução dos scripts

Índice	Campo	Linhas	Valores únicos
1	ANO_BASE	1482960	19
2	CD_PROGRAMA	1482960	5151
3	NM_PROGRAMA	1482960	2554
4	ID_DOCENTE	991187	109877
5	NM_DOCENTE	1482960	145196
6	NM_NACIONALIDADE_DOCENTE	1482960	156

Índice	Campo	Linhas	Valores únicos
7	AN_NASCIMENTO_DOCENTE	1482952	89
8	DS_CATEGORIA_DOCENTE	1482960	3
9	DS_TIPO_VINCULO_DOCENTE	1482941	7
10	NM_ULTIMO_GRAU_TITULACAO	1482946	7
11	AN_ULTIMO_GRAU_TITULACAO	1482946	83
12	SG_IES_ULTIMO_GRAU_TITULACAO	1469108	7011
13	NM_IES_GRAU_TITULACAO	1482946	14748
14	NM_PAIS_ULTIMO_GRAU_TITULACAO	1482946	109
15	CD_AREA_BASICA_TITULACAO	991173	2495
16	NM_AREA_BASICA_TITULACAO	1482946	1197

Fonte: Dados de pesquisa (2024).

A avaliação do DF5_docentes_BR indica que os campos relacionados ao programa — nome e código — possuem a mesma quantidade de linhas que o ano-base. No entanto, há uma discrepância na contagem de valores únicos, confirmando a ocorrência de múltiplos códigos associados a um único nome de programa, conforme identificado na análise dos dados dos programas. Essa variação reforça a necessidade de utilizar o código do programa como referência para integração com outros DataFrames. Além disso, os campos de identificação e nome dos docentes exigem ajustes, pois há registros sem um ID correspondente, o que requer correção para garantir a precisão e integridade do DF5_docentes_BR.

Outras variáveis também apresentam discrepâncias em relação ao total de 1.482.960 linhas. O campo AN_NASCIMENTO_DOCENTE contém 1.482.952 registros, DS_TIPO_VINCULO_DOCENTE possui 1.482.941, enquanto NM_ULTIMO_GRAU_TITULACAO e AN_ULTIMO_GRAU_TITULACAO registram 1.482.946 linhas cada. O campo SG_IES_ULTIMO_GRAU_TITULACAO apresenta uma diferença maior, com 1.469.108 registros, em comparação com NM_IES_GRAU_TITULACAO, que possui 1.482.946 registros. A maior discrepância ocorre no campo CD_AREA_BASICA_TITULACAO, com 991.173 registros, enquanto NM_AREA_BASICA_TITULACAO apresenta 1.482.946, indicando lacunas que necessitam de investigação.

Essa análise inicial aponta para a necessidade de revisar todas as colunas, especialmente porque a maioria dos campos é do tipo caracter, o que pode resultar em inconsistências, como a presença de caracteres não informativos ("-", ";", espaços em branco)

em locais destinados a dados específicos. Portanto, é preciso validar e normalizar esses campos para garantir a integridade do DF5_docentes_BR, assegurando uma base confiável para análises futuras dos perfis docentes nos programas de pós-graduação.

Ajustes do conjunto de dados dos Docentes

Na análise inicial dos dados do DF5_docentes_BR, referentes aos docentes dos programas de pós-graduação *stricto sensu*, a prioridade foi a avaliação dos campos de identificação (ID_DOCENTE) e nome (NM_DOCENTE). Identificou-se que 491.773 registros de nomes de docentes não possuem um código de identificação associado. Para quantificar o esforço necessário na correção desses campos, foi desenvolvido um *script* apresentado no Quadro 57.

Quadro 57 – Script desenvolvido para inserção dos dados dos arquivos de 2013 a 2022 no DF5_docentes_BR

```
import pandas as pd
file_path = '/content/drive/MyDrive/CAPES/datasets/DF5/DF5.csv'
df = pd.read_csv(file_path, encoding='utf-8', sep=';')
# Agrupa as informações pelo campo NM_DOCENTE
grupos = df.groupby('NM_DOCENTE')
# Prepara um DataFrame para armazenar os resultados
resultado = pd.DataFrame(columns=['NM_DOCENTE', 'ID_DOCENTE Existente', 'Campos Nulos'])
for nome, grupo in grupos:
    # Verifica se existe um ID_DOCENTE para este NM_DOCENTE
    id_existente = grupo['ID_DOCENTE'].notnull().any()
    # Verifica se existem campos nulos para este NM_DOCENTE
    campos_nulos = grupo.isnull().any(axis=1).any()
    # Adiciona os resultados ao DataFrame de resultados
    resultado = resultado.append({'NM_DOCENTE': nome,
                                'ID_DOCENTE Existente': id_existente,
                                'Campos_Nulos': campos_nulos}, ignore_index=True)
# Salva o DataFrame de resultados em um arquivo CSV
resultado.to_csv('/content/drive/MyDrive/CAPES/datasets/DF5/resultado.csv', sep=';', encoding='utf-8', index=False)
```

Fonte: Dados de pesquisa (2024).

O script do Quadro 57 estrutura o DF5_docentes_BR com base no campo NM_DOCENTE e, para cada grupo de docentes com o mesmo nome, verifica duas condições: a) presença de pelo menos um ID_DOCENTE válido, indicando que o docente já recebeu um identificador; b) existência de valores nulos dentro do grupo, sinalizando lacunas nos dados que precisam ser corrigidas. Essa abordagem permite avaliar a extensão do problema e definir estratégias para a atribuição de identificadores, garantindo a consistência dos registros no DF5_docentes_BR.

Os resultados dessas verificações são compilados no arquivo 'resultado.csv', registrando, para cada docente, a existência de um ID e a presença de campos nulos. Devido ao volume de aproximadamente 1,5 milhão de registros, os resultados foram exportados para

um arquivo CSV, facilitando a visualização e o refinamento dos dados. Esse procedimento estrutura a base de dados para futuras análises nos programas de pós-graduação.

Após a análise dos resultados e a revisão dos dados na planilha, identificou-se a necessidade de incluir um campo temporário no DF5_docentes_BR, contendo os nomes dos docentes sem acentuação e cedilhas. Isso foi necessário porque nomes aparentemente idênticos correspondiam a indivíduos distintos, diferenciados apenas por esses caracteres. A normalização dos nomes aprimora a identificação e integração dos dados. Para esse ajuste, foi desenvolvido o script apresentado no Quadro 58, que remove acentos e cedilhas, padronizando os registros e consolidando as informações no DF5_docentes_BR.

Quadro 58 – Script desenvolvido para criação de coluna temporária e adição dos nomes dos docentes sem acentos na mesma

```
!pip install unicode
import pandas as pd
import unicode
# Caminho para o arquivo CSV
caminho_arquivo = '/content/drive/MyDrive/CAPES/datasets/DF5/DF5.csv'
# Carregar o DataFrame
df = pd.read_csv(caminho_arquivo, sep=';', encoding='utf-8')
# Função para remover acentos e cedilha
def remover_acentos(texto):
    return unicode.unidecode(texto)
# Criar a nova coluna 'NM_DOCENTE S_ACENTO' aplicando a função para remover acentos
df['NM_DOCENTE S_ACENTO'] = df['NM_DOCENTE'].apply(remover_acentos)
# Reordenar as colunas para colocar 'NM_DOCENTE S_ACENTO' logo após 'NM_DOCENTE'
colunas = df.columns.tolist()
indice_nm_docente = colunas.index('NM_DOCENTE')
colunas_nova_ordem = colunas[:indice_nm_docente+1] + ['NM_DOCENTE S_ACENTO'] + colunas[indice_nm_docente+1:-1]
df = df[colunas_nova_ordem]
# Salvar o DataFrame atualizado no mesmo arquivo
df.to_csv(caminho_arquivo, sep=';', encoding='utf-8', index=False)
```

Fonte: Dados de pesquisa (2024).

A implementação desse *script* permite a busca eficiente pelos nomes dos docentes, facilitando a identificação de casos em que diferentes docentes possuem o mesmo nome, mas diferem em atributos como ID_DOCENTE, nacionalidade ou data de nascimento. O agrupamento dos registros pelo campo NM_DOCENTE_S_ACENTO auxilia na distinção entre homônimos e na correta associação das informações a cada docente. Após o agrupamento, o script exporta os dados para um arquivo CSV, permitindo sua análise em formato de planilha, o que facilita a revisão e a identificação de possíveis inconsistências nos registros. Em seguida, é feito o agrupamento dos docentes, conforme mostra o Quadro 59.

Quadro 59 – Script desenvolvido para exportação dos nomes dos docentes agrupados pela coluna que não contém acentuação

```
import pandas as pd
# Caminho do arquivo DF5.csv
df5_path = "/content/drive/MyDrive/CAPES/datasets/DF5/DF5.csv"
```

```
# Ler o arquivo DF5.csv
df5 = pd.read_csv(df5_path, sep=";", encoding="utf-8")
# Agrupar pela coluna NM_DOCENTE_S_ACENTO e contar a frequência
grouped = df5.groupby('NM_DOCENTE_S_ACENTO').size().reset_index(name='Frequencia')
# Juntar com o DataFrame original para obter as colunas de interesse, incluindo ID_DOCENTE
merged = pd.merge(grouped, df5[['NM_DOCENTE_S_ACENTO', 'ID_DOCENTE', 'NM_NACIONALIDADE_DOCENTE',
'AN_NASCIMENTO_DOCENTE']], on='NM_DOCENTE_S_ACENTO', how='left')
# Remover duplicatas
merged = merged.drop_duplicates()
# Salvar o resultado em um novo arquivo CSV
output_path = "/content/drive/MyDrive/CAPES/datasets/DF5/Resultado_docentes.csv"
merged.to_csv(output_path, index=False, sep=";", encoding="utf-8")
print(f"Resultado salvo em {output_path}")
```

Fonte: Dados de pesquisa (2024).

Esse *script* faz o agrupamento que permite visualizar os dados de forma organizada, facilitando a identificação de inconsistências e ajustes necessários. A execução do script também revelou 125 casos com espaços duplos nos nomes dos docentes, resultando em registros duplicados para um mesmo ID.

Para a solução desse problema, o script descrito no Quadro 60 remove espaços duplos entre *strings* de busca de um campo específico, padronizando os nomes dos docentes no DF5_docentes_BR. O DataFrame corrigido é salvo no mesmo arquivo, preservando a formatação ajustada.

Quadro 60 – Script desenvolvido para remoção de espaços duplos entre strings de um campo determinado previamente

```
import pandas as pd
# Função para remover espaços duplos em uma coluna específica
def remover_espacos_duplos(caminho_arquivo, nome_coluna):
    # Ler o arquivo CSV
    df = pd.read_csv(caminho_arquivo, sep=";", encoding="utf-8")
    # Remover espaços duplos na coluna especificada
    df[nome_coluna] = df[nome_coluna].str.replace(" ", "", regex=False)
    # Salvar o DataFrame atualizado no mesmo arquivo CSV
    df.to_csv(caminho_arquivo, index=False, sep=";", encoding="utf-8")
    print(f"Espaços duplos removidos com sucesso na coluna '{nome_coluna}'.")
# Caminho do arquivo CSV
df5_path = "/content/drive/MyDrive/CAPES/datasets/DF5/DF5.csv"
# Solicitar ao usuário o nome da coluna
nome_coluna = input("Digite o nome da coluna para remover espaços duplos: ")
# Chamar a função de remoção de espaços duplos com os valores fornecidos
remover_espacos_duplos(df5_path, nome_coluna)
```

Fonte: Dados de pesquisa (2024).

Essa padronização feita com o *script* do Quadro 60 assegura a identificação única dos docentes e mantém a consistência dos dados no DF5_docentes_BR, permitindo análises mais precisas e confiáveis. Uma análise na planilha gerada pelo script identificou registros de docentes sem identificação numérica (ID). Para corrigir essa ausência, foi desenvolvido um *script*, descrito no Quadro 61, que cria um dicionário de IDs únicos, associando docentes já identificados àqueles sem ID.

Quadro 61 – Script desenvolvido para inserção de ID aos nomes identificados na coluna NM_DOCENTE_S_ACENTO e sem ID

```
import pandas as pd
# Caminho do arquivo DF5.csv
df5_path = "/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF5/DF5.csv"
# Ler o arquivo DF5.csv
df5 = pd.read_csv(df5_path, sep=";", encoding="utf-8")
# Identificar os IDs únicos para cada NM_DOCENTE_S_ACENTO
ids_unicos = df5.dropna(subset=['ID_DOCENTE']).drop_duplicates('NM_DOCENTE_S_ACENTO')[['NM_DOCENTE_S_ACENTO',
'ID_DOCENTE']].set_index('NM_DOCENTE_S_ACENTO')['ID_DOCENTE']
# Função para preencher ID_DOCENTE faltante
def preencher_id(row):
    if pd.isna(row['ID_DOCENTE']):
        return ids_unicos.get(row['NM_DOCENTE_S_ACENTO'], pd.NA)
    else:
        return row['ID_DOCENTE']
# Aplicar a função para preencher os IDs faltantes
df5['ID_DOCENTE'] = df5.apply(preencher_id, axis=1)
# Salvar o DataFrame atualizado
df5.to_csv(df5_path, index=False, sep=";", encoding="utf-8")
print(f"DataFrame atualizado salvo em {df5_path}")
```

Fonte: Dados de pesquisa (2024).

Nesse *script*, do Quadro 61, foi implementada a função `preencher_id` para verificar campos nulos e, quando possível, atribuir um ID a registros correspondentes. Esse procedimento garante a completude e a coerência do DF5_docentes_BR, preparando os dados para análises futuras. Após a aplicação desse script, foi gerado um novo arquivo CSV atualizado para análise em formato de planilha. A análise na nova planilha revelou particularidades nos registros dos docentes, que não são decorrentes de espaçamento ou pontuação. Identificou-se que alguns docentes aparecem com variações nominais, possivelmente por mudanças no estado civil, como a adoção de um sobrenome de casado em relação ao nome de solteiro, conforme apresentado na Tabela 23.

Tabela 23 – Exemplo de dados encontrados após a execução do script que agrupa os IDs aos nomes

Nome docente	Frequência	ID	Nacionalidade	Ano nascimento
ZULEICA DANTAS PEREIRA	13	537231	BRASIL	1965
ZULEICA DANTAS PEREIRA CAMPOS	6	537231	BRASIL	1965

Fonte: Dados de pesquisa (2024).

O exposto na Tabela 23 evidencia a necessidade de estratégias para unificar registros e garantir a identificação correta dos docentes no DF5_docentes_BR, preservando a consistência das informações.

Para identificar variações nos nomes dos docentes que compartilham o mesmo ID, nacionalidade e ano de nascimento, foi desenvolvido um *script*, descrito no Quadro 62. Esse *script* filtra e analisa casos em que um mesmo docente aparece com diferentes registros de

nome no DF5_docentes_BR. O objetivo é corrigir essas inconsistências e padronizar os registros, assegurando a integridade e a consistência dos dados.

Quadro 62 – Script desenvolvido para filtrar nomes diferentes mas com o mesmo ID

```
import pandas as pd
# Caminho do arquivo DF5.csv
df5_path = "/content/drive/MyDrive/CAPES/datasets/DF5/DF5.csv"
# Ler o arquivo DF5.csv
df5 = pd.read_csv(df5_path, sep=";", encoding="utf-8")
# Filtrar apenas as linhas com ID_DOCENTE preenchido
df5_com_id = df5.dropna(subset=['ID_DOCENTE'])
# Contar a frequência de cada combinação de ID_DOCENTE e NM_DOCENTE_S_ACENTO
frequencias = df5_com_id.groupby(['ID_DOCENTE', 'NM_DOCENTE_S_ACENTO']).size().reset_index(name='frequencia')
# Juntar as frequências com as informações de nacionalidade e ano de nascimento
resultado = pd.merge(frequencias, df5_com_id[['ID_DOCENTE', 'NM_DOCENTE_S_ACENTO', 'NM_NACIONALIDADE_DOCENTE', 'AN_NASCIMENTO_DOCENTE']], drop_duplicates(), on=['ID_DOCENTE', 'NM_DOCENTE_S_ACENTO'], how='left')
# Filtrar para manter apenas os IDs com mais de um nome associado
ids_com_multiplos_nomes = resultado['ID_DOCENTE'].value_counts()[resultado['ID_DOCENTE'].value_counts() > 1].index
resultado_final = resultado[resultado['ID_DOCENTE'].isin(ids_com_multiplos_nomes)]
# Salvar os resultados em um novo arquivo CSV para análise
output_path = "/content/drive/MyDrive/CAPES/datasets/DF5/Docentes_IDs_Multiplos_Nomes_Frequencia.csv"
resultado_final.to_csv(output_path, index=False, sep=";", encoding="utf-8")
print(f"Docentes com mesmo ID, mas nomes diferentes e suas frequências, salvos em {output_path}")
```

Fonte: Dados de pesquisa (2024).

A aplicação do script identificou 43.659 registros com variações nos nomes dos docentes. A inspeção no LibreOffice Calc revelou diferenças sutis, como a inclusão ou omissão de apóstrofes, exemplificado por ROBERTO DALL AGNOL e ROBERTO DALL'AGNOL, indicando a necessidade de ajustes no algoritmo de remoção de acentos para considerar esses caracteres. Além disso, foram encontrados casos de múltiplas nacionalidades, como RAINER RANDOLPH, registrado como brasileiro e alemão, e variações ortográficas em sobrenomes, como BERNADETTE DORA GOBOSSY DE MELO FRANCO e DORA GOMBOSSY DE MELO FRANCO, sugerindo erros de digitação que exigem verificação individual.

Dado o desafio de unificar essas variações com métodos automatizados, como a distância de Levenshtein, optou-se por manter a coluna NM_DOCENTE_S_ACENTO no DF5_docentes_BR, facilitando buscas, enquanto NM_DOCENTE preserva as variações nominais. Entretanto, persistem desafios na normalização automática, como as diferenças entre ARLINDO PHILIPPI JR e ARLINDO PHILIPPI JUNIOR. A revisão do arquivo CSV identificou mais de 5.000 casos semelhantes, exigindo, em alguns casos, consulta ao currículo Lattes para correção.

Para facilitar a atualização manual dos registros, foi desenvolvido um *script*, apresentado no Quadro 63, que permite inserir IDs dos docentes e nomes corrigidos. Dada a complexidade e variabilidade dos dados, essa abordagem manual se mostra necessária para garantir a padronização e a integridade dos registros acadêmicos.

Quadro 63 – Script desenvolvido alterar os nomes dos docentes pelo ID

```
import pandas as pd
# Carregar o DataFrame
file_path = '/content/drive/MyDrive/CAPES/datasets/DF5/DF5.csv'
df = pd.read_csv(file_path, encoding="utf-8", sep=";")
# Dicionário com ID_DOCENTE como chave e NM_DOCENTE_S_ACENTO como valor
# Exemplo: {123: "Nome Sem Acento 1", 456: "Nome Sem Acento 2"}
updates = {
    123: "Nome Sem Acento 1", 456: "Nome Sem Acento 2",
    # Adicione mais IDs e nomes conforme necessário
}
# Atualizar o DataFrame
for id_docente, nm_docente_s_acento in updates.items():
    df.loc[df['ID_DOCENTE'] == id_docente, 'NM_DOCENTE_S_ACENTO'] = nm_docente_s_acento
# Salvar o arquivo atualizado
df.to_csv(file_path, index=False, sep=';', encoding='utf-8')
print("Arquivo atualizado e salvo com sucesso.")
```

Fonte: Dados de pesquisa (2024).

No refinamento do DF5_docentes_BR, que consolida informações sobre docentes em programas de pós-graduação, é implementada uma correção nominal por meio de um *script*. Inicialmente, o conjunto de dados continha 126.097 identificações únicas na coluna NM_DOCENTE_S_ACENTO. Após a unificação de variações nos nomes sem acentuação, esse número foi reduzido para 123.438, eliminando redundâncias.

Durante a análise, foram identificadas fusões indevidas de registros de docentes distintos, como ANDRE LUIS DE OLIVEIRA e ANDRE LUIZ DE OLIVEIRA. A verificação nos currículos Lattes confirmou que possuem IDs e datas de nascimento diferentes, indicando que são homônimos. Essa revisão permitiu corrigir a sobreposição e garantir a precisão dos registros no DF5_docentes_BR, evidenciando a necessidade de verificações detalhadas no tratamento de dados acadêmicos.

Também é realizada uma análise para detectar inconsistências nas datas de nascimento associadas a um mesmo ID_DOCENTE, conforme ilustrado do Quadro 64. Os registros são agrupados pelo identificador do docente e pelo ano de nascimento, revelando 1.192 casos com múltiplas datas. Exemplos incluem IGOR AGOSTINI, com registros para os anos 1973 e 2000, e CARLOS ANTONIO GARRIGA ACOSTA, com registros em 1900 e 1958. Essas inconsistências exigem revisão detalhada para garantir a integridade dos dados no DF5_docentes_BR.

Quadro 64 – Script desenvolvido para filtrar ID e data de nascimento

```
import pandas as pd
# Caminho do arquivo
file_path = '/content/drive/MyDrive/CAPES/datasets/DF5/DF5.csv'
# Ler o arquivo
df = pd.read_csv(file_path, encoding="utf-8", sep=";")
# Filtrar os dados para incluir apenas as colunas de interesse
df_filtered = df[['ID_DOCENTE', 'AN_NASCIMENTO_DOCENTE', 'NM_DOCENTE', 'NM_DOCENTE_S_ACENTO',
```

```
'NM NACIONALIDADE DOCENTE']]
# Agrupar os dados pelo ID_DOCENTE e AN_NASCIMENTO_DOCENTE e contar o número de registros
grouped = df_filtered.groupby(['ID_DOCENTE', 'AN_NASCIMENTO_DOCENTE']).size().reset_index(name='Frequencia')
# Filtrar os grupos para incluir apenas aqueles com mais de uma data de nascimento associada ao mesmo ID_DOCENTE
ids_multiple_birthdates = grouped[~grouped.duplicated('ID_DOCENTE', keep=False)]
# Filtrar o DataFrame original para incluir apenas os registros com IDs identificados anteriormente
df_multiple_birthdates = df_filtered[df_filtered['ID_DOCENTE'].isin(ids_multiple_birthdates)].sort_values(by=['ID_DOCENTE',
'AN_NASCIMENTO_DOCENTE'])
# Juntar o DataFrame filtrado com a contagem de frequência
df_final = pd.merge(df_multiple_birthdates, grouped, on=['ID_DOCENTE', 'AN_NASCIMENTO_DOCENTE'], how='left')
# Remover duplicatas para facilitar a análise, mantendo apenas a primeira ocorrência de cada combinação única
df_final_unique = df_final.drop_duplicates(subset=['ID_DOCENTE', 'AN_NASCIMENTO_DOCENTE'])
# Salvar o resultado em um novo arquivo CSV
output_path = '/content/drive/MyDrive/CAPES/datasets/DF5/DF5_filtered_birthdates_frequency.csv'
df_final_unique.to_csv(output_path, index=False, sep=',', encoding='utf-8')
print(f'Arquivo salvo em: {output_path}')
```

Fonte: Dados de pesquisa (2024).

A análise das discrepâncias revelou a necessidade de revisar individualmente os casos em que as datas de nascimento apresentavam inconsistências, pois essas informações não estão disponíveis em plataformas como o currículo Lattes. Foram identificadas 1.206 ocorrências que exigem correção manual, evitando ajustes automáticos que poderiam introduzir erros. No caso de CARLOS ANTONIO GARRIGA ACOSTA, por exemplo, embora a data mais frequente fosse 1900, a opção escolhida foi 1958, considerando critérios de plausibilidade. Esse exemplo ilustra a complexidade do processo e a importância de uma revisão cuidadosa dos dados.

Para corrigir essas inconsistências, foi desenvolvido um *script*, detalhado na Quadro 65, que utiliza um dicionário para associar ID_DOCENTE às respectivas datas de nascimento (AN_NASCIMENTO_DOCENTE). O *script* percorre esse dicionário, atualizando os valores no DF5_docentes_BR e salvando o resultado em um novo arquivo CSV. Essa abordagem permite revisar as alterações antes de aplicá-las definitivamente ao DF5_docentes_BR, garantindo maior controle e precisão na correção dos dados.

Quadro 65 – Script desenvolvido para ajustar o ano de nascimento para o ID

```
import pandas as pd
# Caminho do arquivo
file_path = '/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF5/DF5.csv'
# Ler o arquivo
df = pd.read_csv(file_path, encoding='utf-8', sep=',')
# Dicionário com ID_DOCENTE como chave e AN_NASCIMENTO_DOCENTE como valor
id_nascimento = {
    ID: ano de nascimento,
    # Adicione mais IDs e anos conforme necessário
}
# Iterar sobre o dicionário e atualizar AN_NASCIMENTO_DOCENTE para os IDs fornecidos
for id_docente, ano_nascimento in id_nascimento.items():
    df.loc[df['ID_DOCENTE'] == id_docente, 'AN_NASCIMENTO_DOCENTE'] = ano_nascimento
# Salvar o arquivo atualizado
output_path = '/content/drive/MyDrive/CAPES/datasets/DF5/DF5_updated.csv'
df.to_csv(output_path, index=False, sep=',', encoding='utf-8')
print(f'Arquivo atualizado salvo em: {output_path}')
```


Fonte: Dados de pesquisa (2024).

Na análise e tratamento dos dados, um *script* é implementado para ajustar as informações dos docentes no DF5_docentes_BR. O conjunto apresentou 145.080 nomes de docentes e 123.319 códigos de identificação (ID) únicos, indicando a presença de homônimos. A verificação identificou múltiplos registros para nomes comuns, como cinco ocorrências de "Marcos Antonio de Oliveira", quatro de "Paulo Roberto de Oliveira" e três de "Rafael Ferreira da Silva". A validação dessas entradas foi realizada por meio da plataforma Lattes, conferindo o nome, a instituição associada e o ano do último título acadêmico.

Alguns nomes, especialmente aqueles menos frequentes no contexto nacional ou com grafia semelhante a nomes estrangeiros, não foram encontrados na plataforma Lattes. Nesses casos, são realizadas pesquisas adicionais em motores de busca para verificar a vinculação dos indivíduos a instituições acadêmicas. Essa etapa permite ajustes no banco de dados, garantindo a consistência dos registros de docentes no DF5_docentes_BR.

Nesse ponto da análise, a nacionalidade dos docentes ainda não é analisada detalhadamente. Uma revisão preliminar identificou variações na grafia de certas nacionalidades, como "ALBANIA" e "ALBÂNIA" ou "CANADA" e "CANADÁ", além da presença do termo "NÃO INFORMADO", indicando ausência de informação nos registros. Em alguns casos, um mesmo docente foi registrado com nacionalidades diferentes ao longo dos anos, alternando entre "NÃO INFORMADO", "França" e "Brasil", o que gera inconsistências. Como os registros abrangem o período de 2004 a 2022, uma abordagem possível é considerar as informações mais recentes disponíveis, embora essa estratégia não garanta total precisão.

Para tratar essa questão, é desenvolvido o *script* apresentado no Quadro 66, que calcula a frequência de cada nacionalidade com base no ID do docente e na nacionalidade registrada. Os resultados são armazenados em um arquivo CSV, permitindo a análise em planilhas para identificação e correção das inconsistências.

Quadro 66 – Script desenvolvido para calcular a frequência de cada nacionalidade

```
import pandas as pd
# Caminho do arquivo
file_path = '/content/drive/MyDrive/CAPES/datasets/DF5/DF5.csv'
# Ler o arquivo
df = pd.read_csv(file_path, encoding="utf-8", sep=";")
# Filtrar os dados para incluir apenas as colunas de interesse
df_filtered = df[['ID_DOCENTE', 'AN_NASCIMENTO_DOCENTE', 'NM_DOCENTE', 'NM_DOCENTE_S_ACENTO',
'NM_NACIONALIDADE_DOCENTE']]
# Agrupar os dados pelo ID_DOCENTE e AN_NASCIMENTO_DOCENTE e contar o número de registros
grouped = df_filtered.groupby(['ID_DOCENTE', 'NM_NACIONALIDADE_DOCENTE']).size().reset_index(name='Frequencia')
# Filtrar os grupos para incluir apenas aqueles com mais de uma data de nascimento associada ao mesmo ID_DOCENTE
ids_multiple_birthdates = grouped[grouped.duplicated('ID_DOCENTE', keep=False)]
# Filtrar o DataFrame original para incluir apenas os registros com IDs identificados anteriormente
df_multiple_birthdates = df_filtered[df_filtered['ID_DOCENTE'].isin(ids_multiple_birthdates)].sort_values(by=['ID_DOCENTE',
'AN_NASCIMENTO_DOCENTE'])
# Juntar o DataFrame filtrado com a contagem de frequência
df_final = pd.merge(df_multiple_birthdates, grouped, on=['ID_DOCENTE', 'NM_NACIONALIDADE_DOCENTE'], how='left')
# Remover duplicatas para facilitar a análise, mantendo apenas a primeira ocorrência de cada combinação única
df_final_unique = df_final.drop_duplicates(subset=['ID_DOCENTE', 'NM_NACIONALIDADE_DOCENTE'])
# Salvar o resultado em um novo arquivo CSV
output_path = '/content/drive/MyDrive/CAPES/datasets/DF5/DF5_filtered_nacionalidade_frequency.csv'
df_final_unique.to_csv(output_path, index=False, sep=',', encoding='utf-8')
print(f"Arquivo salvo em: {output_path}")
```

Fonte: Dados de pesquisa (2024).

Após a execução do *script*, identificou-se que 1.128 docentes possuem múltiplas nacionalidades cadastradas. A análise revelou inconsistências devido ao formato do campo, que permite a inserção de letras, números e caracteres especiais (\$, %, &). Para corrigir essa questão, foi desenvolvido o *script* apresentado no Quadro 67, que lê os dados do DF5_docentes_BR, consolida as informações e exibe os resultados agregados. Essa etapa assegura a padronização dos dados antes da análise da distribuição nacional dos docentes.

Quadro 67 – Script desenvolvido para analisar os nomes dos países de nascimento dos docentes

```
import pandas as pd
# Caminho do arquivo
file_path = '/content/drive/MyDrive/CAPES/datasets/DF5/DF5.csv'
# Ler o arquivo
df = pd.read_csv(file_path, encoding="utf-8", sep=";")
# Extrair os valores únicos da coluna 'NM_NACIONALIDADE_DOCENTE' e ordená-los
unique_nationalities = sorted(df['NM_NACIONALIDADE_DOCENTE'].unique())
# Salvar as nacionalidades únicas em um arquivo CSV
unique_nationalities_df = pd.DataFrame(unique_nationalities, columns=['Nacionalidade Única'])
unique_nationalities_df.to_csv('nacionalidades_unicas.csv', index=False)
# Calcular a frequência de cada nacionalidade
frequency = df['NM_NACIONALIDADE_DOCENTE'].value_counts().sort_index()
# Salvar a frequência em um arquivo CSV
frequency_df = pd.DataFrame(frequency).reset_index()
frequency_df.columns = ['Nacionalidade', 'Frequência']
frequency_df.to_csv('/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF5/frequencia_nacionalidades.csv', index=False)
print("Os dados foram salvos com sucesso em dois arquivos CSV.")
```

Fonte: Dados de pesquisa (2024).

A aplicação do *script* resultou na identificação de variações na grafia dos nomes de países, como "ALBANIA", "ARMÊNIA", "CORÉIA DO SUL" e "RUSSIA e RÚSSIA",

"CANADA e "CANADÁ". Além disso, algumas entradas não correspondem a países, incluindo "NI" (2 ocorrências), "OUTRO" (141 ocorrências) e "NÃO INFORMADO" (60 ocorrências). Para padronizar alguns dos registros, as entradas que não correspondem a países foram convertidas para "NÃO INFORMADO". Essa padronização foi implementada por meio do *script* apresentado no Quadro 68, que utiliza um dicionário denominado 'replacements', onde as chaves representam as denominações originais e os valores correspondem às correções aplicadas.

Quadro 68 – Script desenvolvido para ajustar os nomes dos países de nascimento dos docentes

```
import pandas as pd
# Caminho do arquivo
file_path = '/content/drive/MyDrive/CAPES/datasets/DF5/DF5.csv'
# Ler o arquivo
df = pd.read_csv(file_path, encoding="utf-8", sep=";")
# Dicionário de substituições
replacements = {
    "ALBANIA": "ALBÂNIA", "AUSTRIA": "ÁUSTRIA", "ARMÉNIA": "ARMÊNIA", "BIELORUSSIA": "BIELORRÚSSIA", "BOLIVIA": "BOLÍVIA", "BULGARIA": "BULGÁ", "CANADA": "CANADÁ", "COLOMBIA": "COLÔMBIA", "COREIA DO SUL": "CORÉIA DO SUL", "NI": "NÃO INFORMADO", "OUTRO": "NÃO INFORMADO", "POLONIA": "POLÔNIA", "CAMARÕES": "REP. DOS CAMARÕES", "RUSSIA": "RÚSSIA", "SUIÇA": "SUÍÇA", "CHECOSLOVAQUIA": "TCHECOSLOVÁQUIA", "REP. DOS CAMARÕES": "REPÚBLICA DOS CAMARÕES"}
# Substituir valores no DataFrame
df["NM_NACIONALIDADE_DOCENTE"] = df["NM_NACIONALIDADE_DOCENTE"].replace(replacements)
# Salvar o DataFrame alterado (opcionalmente em um novo arquivo para evitar a perda de dados)
df.to_csv('/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF5/DF5_updated.csv', index=False, encoding='utf-8', sep=';')
print("Substituições realizadas com sucesso e DataFrame salvo.")
```

Fonte: Dados de pesquisa (2024).

Após a execução do *script*, o DF5_docentes_BR é salvo com as alterações. Foram atualizados 128 países distintos nas nacionalidades dos docentes, além da categoria "NÃO INFORMADO" para registros sem especificação.

Após essa etapa, eventuais ajustes adicionais são aplicados utilizando o *script* apresentado no Quadro 69, que utiliza um dicionário denominado 'updates', no qual as chaves representam os IDs dos docentes e os valores correspondem às nacionalidades atualizadas no DF5_docentes_BR.

Quadro 69 – Script desenvolvido para alterar a nacionalidade dos docentes de acordo com o ID

```
import pandas as pd
# Caminho do arquivo
file_path = '/content/drive/MyDrive/CAPES/datasets/DF5/DF5.csv'
# Ler o arquivo
df = pd.read_csv(file_path, encoding="utf-8", sep=";")
# Dicionário com os IDs dos docentes e as nacionalidades correspondentes
updates = {
    id: "país",
    # Adicione mais IDs e nacionalidades conforme necessário
}
# Atualizar a coluna 'NM_NACIONALIDADE_DOCENTE' com base no dicionário
```

```
for doc_id, nationality in updates.items():
    df.loc[df['ID_DOCENTE'] == doc_id, 'NM_NACIONALIDADE_DOCENTE'] = nationality
# Salvar o DataFrame atualizado em um novo arquivo para evitar sobrescrever o original
df.to_csv('/content/drive/MyDrive/CAPEs/datasets/dados_abertos/DF5/DF5_updated.csv', index=False, encoding='utf-8', sep=';')
print("Atualizações realizadas com sucesso e DataFrame salvo.")
```

Fonte: Dados de pesquisa (2024).

O ajuste do *script* do Quadro 69 padroniza as informações sobre a origem nacional dos docentes catalogados.

Seguindo a análise, passa-se à avaliação do campo referente à categoria dos docentes. Esse campo, do tipo *string*, contém três valores: COLABORADOR (272.315), PERMANENTE (1.189.190) e VISITANTE (21.455). Como essa classificação se baseia em dados internos de cada PPG, e não possui correspondência verificável e confiável em fontes externas, nenhuma alteração é realizada nesse campo. Assim, as classificações originais são mantidas para a análise subsequente.

Em seguida, a análise se concentra no campo que indica o tipo de vínculo dos docentes com suas instituições. A documentação desse campo é limitada, informando apenas o vínculo do docente com a instituição. No DF5_docentes_BR, os valores registrados são: APOSENTADO (35.177), BOLSA DE FIXAÇÃO (11.099), CLT (140.420), COLABORADOR (86.888), NÃO INFORMADO (279), SERVIDOR PÚBLICO (763.765) e SERVIDOR PÚBLICO/CLT (445.313). Também para esse campo há somente dados internos do PPG, não havendo fontes externas para validação, o que impossibilita correções ou adaptações.

A Tabela 22 indica que esse campo contém 1.482.941 registros, com 19 linhas apresentando valores nulos ou em branco. Para identificar essas inconsistências, é desenvolvido o *script* ilustrado no Quadro 70, que localiza e extrai as linhas com valores vazios, salvando-as em um arquivo CSV.

Quadro 70 – Script desenvolvido para filtrar as linhas que estão sem valor para o campo DS_TIPO_VINCULO_DOCENTE

```
import pandas as pd
# Caminho do arquivo
file_path = '/content/drive/MyDrive/CAPEs/datasets/DF5/DF5.csv'
# Ler o arquivo
df = pd.read_csv(file_path, encoding='utf-8', sep=';')
# Filtrar as linhas onde 'NM_ULTIMO_GRAU_TITULACAO' está vazio ou nulo
missing_data = df[df['DS_TIPO_VINCULO_DOCENTE'].isna()]
# Salvar essas linhas em um novo arquivo CSV
missing_data.to_csv('/content/drive/MyDrive/CAPEs/datasets/DF5/missing_tipo_vinculo.csv', index=False, encoding='utf-8', sep=';')
print("Dados com 'DS_TIPO_VINCULO_DOCENTE' vazio ou nulo foram salvos no arquivo CSV.")
```

Fonte: Dados de pesquisa (2024).

Depois de atualizado, o arquivo CSV pode ser aberto em uma planilha para análise detalhada, permitindo a definição das correções necessárias para manter a integridade do conjunto de dados. O *script* apresentado no Quadro 71 identifica se o campo DS_TIPO_VINCULO_DOCENTE está vazio ou nulo. A atualização no DF5_docentes_BR ocorre apenas nesses casos, utilizando valores específicos para cada ID_DOCENTE.

Quadro 71 – Script desenvolvido para inserir valores faltantes no campo DS_TIPO_VINCULO_DOCENTE

```
import pandas as pd
# Caminho do arquivo
file_path = '/content/drive/MyDrive/CAPES/datasets/DF5/DF5.csv'
# Ler o arquivo
df = pd.read_csv(file_path, encoding='utf-8', sep=';')
# Dicionário com os IDs dos docentes e os tipos de vínculo correspondentes
updates = {
    156928: "SERVIDOR PÚBLICO", 50774: "BOLSA DE FIXAÇÃO", 452467: "SERVIDOR PÚBLICO", 16918:
    "APOSENTADO", 514760: "SERVIDOR PÚBLICO", 160242: "SERVIDOR PÚBLICO", 22439: "SERVIDOR PÚBLICO",
    65728: "BOLSA DE FIXAÇÃO", 198632: "NÃO INFORMADO", 185173: "NÃO INFORMADO", 17052: "SERVIDOR
    PÚBLICO", 714173: "SERVIDOR PÚBLICO", 20986: "SERVIDOR PÚBLICO/CLT", 59625: "NÃO INFORMADO",
    490771: "NÃO INFORMADO", 971127: "NÃO INFORMADO", 231678: "NÃO INFORMADO", 17020: "SERVIDOR
    PÚBLICO/CLT", 86188: "NÃO INFORMADO"
}
# Adicione mais IDs e tipos de vínculo conforme necessário
}
# Atualizar 'DS_TIPO_VINCULO_DOCENTE' apenas se estiver vazio ou nulo
for doc_id, vinculo in updates.items():
    # Encontrar linhas onde ID_DOCENTE corresponde e o DS_TIPO_VINCULO_DOCENTE está vazio
    mask = (df['ID_DOCENTE'] == doc_id) & (df['DS_TIPO_VINCULO_DOCENTE'].isna())
    df.loc[mask, 'DS_TIPO_VINCULO_DOCENTE'] = vinculo
# Salvar o DataFrame atualizado em um novo arquivo para evitar sobrescrever o original
df.to_csv('/content/drive/MyDrive/CAPES/datasets/DF5/DF5.csv', index=False, encoding='utf-8', sep=';')
print("Dados atualizados com sucesso e DataFrame salvo.")
```

Fonte: Dados de pesquisa (2024).

O procedimento do Quadro 68 garante que apenas os registros que requerem correção sejam modificados, preservando a consistência das informações no banco de dados.

Em seguida, a análise vai para o campo referente ao último grau acadêmico obtido pelos docentes, que indica o nível mais alto de titulação registrado. De acordo com a documentação, esse campo é fundamental para a caracterização da qualificação dos docentes. A Tabela 24 apresenta os títulos acadêmicos e suas respectivas quantidades, com destaque para o título de Doutorado, que corresponde a 99,30% dos registros.

Tabela 24 – Relação de nomes do último grau de titulação dos docentes

NM_ÚLTIMO_GRAU_TITULACAO	Frequência	%
BACHARELADO	575	0.04%
DOUTORADO	1.472.501	99.30%
DOUTORADO PROFISSIONAL	26	0.00%
GRADUAÇÃO	268	0.02%

NM_ÚLTIMO_GRAU_TITULACAO	Frequência	%
MESTRADO	8437	0.57%
MESTRADO PROFISSIONAL	607	0.04%
NOTÓRIO SABER	532	0.04%
Total	1.482.946	100%

Fonte: Dados de pesquisa (2024).

Analisando esse campo no DF5_docentes_BR, verifica-se que contém 1.482.960 registros, portanto, como o total na tabela é menor, percebe-se uma discrepância de 14 registros. Essa inconsistência requer verificação antes da continuidade das análises. Para tanto, é implementado o script (Quadro 72), que utiliza o método .isna() para identificar registros em que a coluna NM_ULTIMO_GRAU_TITULACAO contenha valores vazios ou nulos.

Quadro 72 – Script desenvolvido para filtrar as linhas que estão sem valor para o campo NM_ULTIMO_GRAU_TITULACAO

```
import pandas as pd
# Caminho do arquivo
file_path = '/content/drive/MyDrive/CAPES/datasets/DF5/DF5.csv'
# Ler o arquivo
df = pd.read_csv(file_path, encoding='utf-8', sep=';')
# Filtrar as linhas onde 'NM_ULTIMO_GRAU_TITULACAO' está vazio ou nulo
missing_data = df[df['NM_ULTIMO_GRAU_TITULACAO'].isna()]
# Salvar essas linhas em um novo arquivo CSV
missing_data.to_csv('/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF5/missing_grau_titulacao.csv', index=False,
encoding='utf-8', sep=';')
print("Dados com 'NM_ULTIMO_GRAU_TITULACAO' vazio ou nulo foram salvos no arquivo CSV.")
```

Fonte: Dados de pesquisa (2024).

Esses registros são filtrados e exportados para um arquivo CSV, permitindo a visualização em planilha para análise detalhada.

Após a execução do script apresentado no Quadro 72, a análise do arquivo gerado identificou 14 linhas com valores nulos. Além do campo NM_ULTIMO_GRAU_TITULACAO vazio, outros campos subsequentes também apresentaram ausência de dados, como Ano do Último Grau de Titulação, Sigla da Instituição de Ensino Superior do Último Grau e Nome da Área Básica da Titulação, conforme mostrado na Figura 7.

Figura 7 – Situação dos campos vazios após a execução do script da Quadro 73

	J	K	L	M	N	O	P	Q
1	DS TIPO VINCULO DOCENTE	NM_ULTIMO_GRAU_TITULACAO	AN_ULTIMO_GRAU_TITULACAO	SG_IES_ULTIMO_GRAU_TITULACAO	NM_IES_GRAU_TITULACAO	NM_PAIS_ULTIMO_GRAU_TITULACAO	CD_AREA_BASICA_TITULACAO	NM_AREA_BASICA_TITULACAO
2	SERVIDOR PUBLICO							
3	SERVIDOR PUBLICO							
4	SERVIDOR PUBLICO							
5	SERVIDOR PUBLICO							
6	CLT							
7	SERVIDOR PUBLICO							
8	SERVIDOR PUBLICO							
9	SERVIDOR PUBLICO							
10	SERVIDOR PUBLICO							
11	SERVIDOR PUBLICO							
12	SERVIDOR PUBLICO							
13	SERVIDOR PUBLICO							
14	CLT							
15	CLT							
16								

Fonte: Dados de pesquisa (2024).

As inconsistências identificadas na Figura 7 apontam lacunas nos registros, exigindo revisão detalhada para garantir a integridade do DF5_docentes_BR. Para tratar essas inconsistências, foi utilizado o *script* ilustrado no Quadro 73, que insere o nome do docente sem acentuação para recuperar todas as linhas correspondentes a esse indivíduo no *dataset*.

Quadro 73 – Script desenvolvido para salvar em um arquivo csv todas as linhas pelo nome do docente sem acento

```
import pandas as pd
# Caminho do arquivo
file_path = '/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF5/DF5.csv'
# Ler o arquivo
df = pd.read_csv(file_path, encoding="utf-8", sep=";")
# Solicitar ao usuário o nome da coluna e o valor a ser filtrado
nome_coluna = "NM_DOCENTE_S_ACENTO"
valor_coluna = "CARLOS ALBERTO PEREZ"
# Filtrar o DataFrame com base no valor da coluna
df_filtrado = df[df[nome_coluna] == valor_coluna]
# Definir o nome do arquivo de saída com base no nome da coluna e no valor
nome_arquivo_saida =
f'/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF5/Script_valida_coluna_{nome_coluna}_campo_{valor_coluna}.csv'
# Salvar o DataFrame filtrado no arquivo de saída
df_filtrado.to_csv(nome_arquivo_saida, index=False, sep=";", encoding="utf-8")
print(f'Arquivo salvo com sucesso: {nome_arquivo_saida}')
```

Fonte: Dados de pesquisa (2024).

A partir dessa consulta, os dados mais relevantes ou próximos ao período das lacunas são selecionados para atualização no DF5_docentes_BR. O resultado é exportado para um arquivo CSV, permitindo a análise em planilha. Essa abordagem facilita a verificação dos registros e orienta a aplicação das correções necessárias no DF5_docentes_BR, garantindo a consistência dos dados.

A execução do *script* do Quadro 73 gerou 14 arquivos CSV. A análise identificou que as omissões ocorreram apenas em 2013, indicando falha no arquivo de origem que afetou registros de diferentes programas ou docentes. Para corrigir, é criado um dicionário (updates), no qual cada chave representa um ID_DOCENTE e os valores correspondem às colunas a

serem atualizadas e seus novos dados. O script do Quadro 74 percorre esse dicionário, localiza os registros no DF5_docentes_BR, aplica as correções e salva as modificações, garantindo a consistência dos dados.

Quadro 74 – Script desenvolvido para inserir valores faltantes no campo DS_TIPO_VINCULO_DOCENTE

```
import pandas as pd
# Caminho do arquivo
file_path = '/content/drive/MyDrive/CAPES/datasets/DF5/DF5.csv'
# Ler o arquivo
df = pd.read_csv(file_path, encoding="utf-8", sep=";")
# Dicionário com os dados a serem atualizados
updates = {19811: {"NM_ULTIMO_GRAU_TITULACAO": "DOUTORADO", "AN_ULTIMO_GRAU_TITULACAO":
2000, "SG_IES_ULTIMO_GRAU_TITULACAO": "ESM", "NM_IES_GRAU_TITULACAO": "UNIVERSITY OF ROCHESTER",
"NM_PAIS_ULTIMO_GRAU_TITULACAO": "ESTADOS UNIDOS", "CD_AREA_BASICA_TITULACAO": 80303005,
"NM_AREA_BASICA_TITULACAO": "MÚSICA" }
# Adicione mais IDs e dados conforme necessário
}
# Atualizar as colunas no DataFrame com base no ID_DOCENTE
for doc_id, data in updates.items():
    index = df[df['ID_DOCENTE'] == doc_id].index
    for column, value in data.items():
        df.loc[index, column] = value
# Salvar o DataFrame atualizado em um novo arquivo para evitar sobrescrever o original
df.to_csv('/content/drive/MyDrive/CAPES/datasets/DF5/DF5.csv', index=False, encoding='utf-8', sep=';')
print("Dados atualizados com sucesso e DataFrame salvo.")
```

Fonte: Dados de pesquisa (2024).

Passando para a análise do campo de ano do último grau de titulação, são identificados registros de 1936 a 2209. Enquanto 1936 pode indicar um nascimento por volta de 1909, 2209 representa um erro. Para corrigir essa inconsistência, utiliza-se o *script* do Quadro 75, que extrai registros com anos anteriores a 1950 ou posteriores a 2022.

Quadro 75 – Script desenvolvido para obter os dados do DF5_docentes_BR que contenham AN_ULTIMO_GRAU_TITULACAO menor que 1950 e maior que 2022

```
import pandas as pd
# Caminho do arquivo
file_path = '/content/drive/MyDrive/CAPES/datasets/DF5/DF5.csv'
# Ler o arquivo
df = pd.read_csv(file_path, encoding="utf-8", sep=";")
# Filtrar dados onde 'AN_ULTIMO_GRAU_TITULACAO' é menor que 1950 ou maior que 2022
filtered_df = df[(df['AN_ULTIMO_GRAU_TITULACAO'] < 1950) | (df['AN_ULTIMO_GRAU_TITULACAO'] > 2022)]
# Salvar os dados filtrados em um novo arquivo CSV
filtered_df.to_csv('/content/drive/MyDrive/CAPES/datasets/filtered_AN_ULTIMO_GRAU_TITULACAO.csv', index=False,
encoding='utf-8', sep=';')
print("Os dados filtrados foram salvos com sucesso.")
```

Fonte: Dados de pesquisa (2024).

Com a execução desse *script*, foram identificados 172 casos, que serão verificados via currículo Lattes e buscadores online para possíveis ajustes. A análise identificou registros em que o ano de nascimento coincidia com o da última titulação ou era posterior a ele, indicando erro. Também foram encontrados casos corretamente registrados no DF5_docentes_BR. Para

corrigir as inconsistências, 16 registros foram atualizados por meio do *script* do Quadro 76, que ajusta o campo AN_ULTIMO_GRAU_TITULACAO no DF5_docentes_BR com base em um conjunto específico de IDs, garantindo a consistência dos dados.

Quadro 76 – Script desenvolvido para atualizar o campo AN_ULTIMO_GRAU_TITULACAO

```
import pandas as pd
# Caminho do arquivo
file_path = '/content/drive/MyDrive/CAPES/datasets/DF5/DF5.csv'
# Ler o arquivo
df = pd.read_csv(file_path, encoding="utf-8", sep=";")
# Dicionário com os IDs dos docentes e os novos anos de titulação correspondentes
updates = {19571: 1978, 12782: 1995, 527944: 1996, 18203: 1988, 41: 1998, 8003928: 1995, 18374: 1976, 8005580: 1974, 8005743: 2008, 529320: 2000, 271810: 2009, 8007911: 1956, 10229: 1960, 17407: 1960, 18469: 1977, 8011605: 1970, 8013060: 1970, 8013165: 1970, 8013584: 1946}
# Adicione mais IDs e anos conforme necessário
}
# Atualizar 'AN_ULTIMO_GRAU_TITULACAO' com base no ID_DOCENTE
for doc_id, year in updates.items():
    df.loc[df['ID_DOCENTE'] == doc_id, 'AN_ULTIMO_GRAU_TITULACAO'] = year
# Salvar o DataFrame atualizado em um novo arquivo para evitar sobrescrever o original
df.to_csv('/content/drive/MyDrive/CAPES/datasets/DF5/DF5.csv', index=False, encoding='utf-8', sep=';')
print("Atualizações realizadas com sucesso e DataFrame salvo.")
```

Fonte: Dados de pesquisa (2024).

A seguir, a análise dos campos SG_IES_ULTIMO_GRAU_TITULACAO (siglas) e NM_IES_GRAU_TITULACAO (nomes) identificou discrepâncias nos registros. O campo de siglas contém 1.469.231 registros e 6.991 valores únicos, enquanto o de nomes possui 1.482.960 registros e 14.706 valores únicos, indicando que 13.729 linhas apresentam nomes sem siglas. Além disso, a variação no número de valores únicos sugere inconsistências, pois cada sigla deveria estar associada a uma única IES.

Para tratar essas questões, o *script* descrito no Quadro 77 extrai conjuntos únicos de siglas, nomes e países (NM_PAIS_ULTIMO_GRAU_TITULACAO) e exporta os dados para um arquivo CSV, permitindo a análise e correção das entradas para alinhar siglas e nomes das instituições.

Quadro 77 – Script desenvolvido para extrair as informações dos campos SG_IES_ULTIMO_GRAU_TITULACAO e NM_IES_GRAU_TITULACAO

```
import pandas as pd
# Caminho do arquivo
file_path = '/content/drive/MyDrive/CAPES/datasets/DF5/DF5.csv'
# Ler o arquivo
df = pd.read_csv(file_path, encoding="utf-8", sep=";")
# Agrupar por nome, sigla e país das instituições e contar frequências
institution_counts = df.groupby(['NM_IES_GRAU_TITULACAO', 'SG_IES_ULTIMO_GRAU_TITULACAO', 'NM_PAIS_ULTIMO_GRAU_TITULACAO']).size().reset_index(name='Frequencia')
# Salvar os resultados em um novo arquivo CSV
institution_counts.to_csv('/content/drive/MyDrive/CAPES/datasets/DF5/institution_frequencies.csv', index=False, encoding='utf-8', sep=';')
print("Frequências das instituições foram calculadas e salvas com sucesso.")
```

Fonte: Dados de pesquisa (2024).

A execução do *script* gerou uma planilha com 23.387 linhas, identificando registros inconsistentes, conforme ilustrado na Figura 8.

Figura 8 – Visualização de dados referentes ao nome da IES da última titulação

NM IES GRAU_TITULACAO	SG IES_ULTIMO_GRAU_TITULACAO	NM PAIS_ULTIMO_GRAU_TITULACAO	Frequencia
111	NÃO INFORMADA	NÃO INFORMADO	2
11111111	NÃO INFORMADA	ESTADOS UNIDOS	1
-UNIVERSIDADE DE SÃO PAULO	USP	BRASIL	1
PONTIFÍCIA UNIVERSIDADE CATÓLICA DE SÃO PAULO - PUC-SP	PUC-SP	BRASIL	4
UNIVERSIDADE ESTADUAL DE CAMPINAS	UNICAMP	BRASIL	3
	CIRE	FRANÇA	3
	NÃO INFORMADA	NÃO INFORMADO	1
	NÃO INFORMADA	NÃO INFORMADO	1
UFMT	UFMT	BRASIL	1
UNIVERSIDADE FEDERAL DE VIÇOSA	UFV	BRASIL	7
[EL] ACCORD ECOLE DE LANGUES	ACCORD	FRANÇA	8
[EL] CIREFE - UNIVERSITÉ DE RENNES 2	NI	FRANÇA	1
[EL] INSA LYON - SERVICE DE FRANÇAIS LANGUE ÉTRANGÈRE (SFE)	NI	FRANÇA	6
[EL] S + W SPEAK + WRITE	NI	ALEMANHA	6
[EL] UNIVERSITÉ DE RENNES 2 - CIREFE	NI	FRANÇA	18
[EL] DID DEUTSCH-INSTITUT, FRANKFURT	DID	ALEMANHA	4
[EL] FRENCH IN NORMANDY	NI	FRANÇA	1
[EL] INSTITUT DE FRANÇAIS - UNIVERSITÉ D'ORLÉANS, ORLÉANS	NI	FRANÇA	6
[UE] ALBERT EINSTEIN COLLEGE OF MEDICINE	NI	ESTADOS UNIDOS	49
[UE] CENTRO PARA EL DESARROLLO DE LA INVESTIGACIÓN CIENTÍFICA	CEDIC	PARAGUAI	11
[UE] ENEDIS DR LORRAINE	NI	FRANÇA	15
[UE] FACULDADE DE PSICOLOGIA E DE CIÊNCIAS DA EDUCAÇÃO	FPCEUP	PORTUGAL	1
[UE] FACULTAD DE CIENCIAS ECONÓMICAS Y EMPRESARIALES	NI	ESPAÑA	3
[UE] INSTITUTO NACIONAL DE SAÚDE PÚBLICA	INSP	CABO VERDE	4
[UE] SANOFI - AVENTIS DEUTSCHLAND GMBH, FRANKFURT	SANOFI	ALEMANHA	4
UNIVERSIDADE FEDERAL DE VIÇOSA	UFV	BRASIL	1
UNIVERSIDADE FEDERAL DO RIO DE JANEIRO	UFRJ	BRASIL	2
A	NÃO INFORMADA	NÃO INFORMADO	2
A T STILL UNIVERSITY OF HEALTH SCIENCES	ATSU	ESTADOS UNIDOS	2
ÉCOLE NATIONALE DES PONTS ET CHAUSSEES	ENPC	FRANÇA	1
A.E. OFFE PHYSICO-TECHNICAL INSTITUTE	ISPETERSBURG	RUSSIA	1

Fonte: Dados de pesquisa (2024).

A análise identificou erros de inserção e digitação, como nomes de instituições registrados como '111', '11111111', 'INSTITUIÇÃO NÃO CADASTRADA' e 'KSDIJFWOIJHEFROIWEROIW'. Siglas foram preenchidas incorretamente, incluindo 'NÃO INFORMADA', 'UFMT' e 'UNIVERSIDADE NOS ESTADOS UNIDOS'. Também foram encontradas inconsistências, como o uso do caractere '.' para representar o nome da instituição e a designação "NÃO INFORMADA" para a sigla. Além disso, foram detectados caracteres especiais não pertencentes ao alfabeto brasileiro, como 'Ł' e 'ø', e abreviações irregulares, comprometendo a identificação das instituições. Variações na nomenclatura, como 'PARIS 1' e 'PARIS I', também foram observadas. Por fim, registros de faculdades ou institutos como entidades separadas levaram à decisão de agrupá-los sob o nome principal da universidade.

Para resolver as inconsistências identificadas, foram realizadas verificações cruzadas no currículo Lattes, buscadores online e sites institucionais para corrigir as denominações das instituições onde os orientadores concluíram seus estudos, garantindo a precisão dos registros no DF5_docentes_BR. O conjunto de dados consolidado do DF5_docentes_BR contém 2.162 instituições, com 1.980 siglas distintas. Contudo, o campo de siglas não foi utilizado nas

análises posteriores devido a problemas como nomes idênticos para IES em diferentes países e registros imprecisos marcados como "NI". No campo referente ao país da última titulação, foram identificados 78 países, e casos não confirmados permaneceram com a marcação "NÃO INFORMADA".

A análise dos códigos e nomes das áreas básicas da última titulação revelou discrepâncias: 991.173 registros para o código e 1.482.946 para o nome, com 2.495 códigos únicos e 1.197 nomes únicos. Conforme a Capes, a estrutura das Áreas do Conhecimento segue quatro níveis hierárquicos: Grande Área (1º nível), Área do Conhecimento (2º nível), Subárea (3º nível) e Especialidade (4º nível), sendo que, no DF5_docentes_BR, os registros devem corresponder ao 2º nível.

Para verificar e corrigir essas inconsistências, foi desenvolvido um *script*, exposto no Quadro 78, que compila e exporta os nomes e códigos das áreas de titulação para um arquivo CSV, permitindo a análise e validação dos dados em planilhas.

Quadro 78 – Script desenvolvido para recuperar os nomes das áreas de titulação básica e seus códigos

```
import pandas as pd
# Caminho do arquivo
file_path = '/content/drive/MyDrive/CAPES/datasets/DF5/DF5.csv'
# Ler o arquivo
df = pd.read_csv(file_path, encoding="utf-8", sep=";")
# Assegurar que todos os valores nulos em CD_AREA_BASICA_TITULACAO sejam tratados como strings vazias
df['CD_AREA_BASICA_TITULACAO'] = df['CD_AREA_BASICA_TITULACAO'].fillna("")
# Selecionar as colunas de interesse
result_df = df[['NM_AREA_BASICA_TITULACAO', 'CD_AREA_BASICA_TITULACAO']]
# Eliminar duplicatas, mantendo a primeira ocorrência
result_df = result_df.drop_duplicates().reset_index(drop=True)
# Salvar o DataFrame resultante em um novo arquivo CSV
output_file_path = '/content/drive/MyDrive/CAPES/datasets/DF5/NM_CD_AREA_BASICA_TITULACAO.csv'
result_df.to_csv(output_file_path, index=False, encoding='utf-8', sep=';')
print("Arquivo com áreas básicas e seus códigos foi salvo com sucesso em:", output_file_path)
```

Fonte: Dados de pesquisa (2024).

A execução desse *script* gerou um arquivo CSV com 3.281 linhas, analisado em formato de planilha. A revisão indicou que os registros seguiam três níveis hierárquicos das Áreas do Conhecimento definidas pela Capes. Inicialmente, o foco foi a padronização das áreas específicas, mas a análise dos currículos Lattes dos orientadores revelou que muitos registros, especialmente os associados à área "CIÊNCIAS EXATAS E DA TERRA", continham apenas a Grande Área (1º nível), sem detalhamento das áreas subsequentes.

Para acomodar essa estrutura, foram adicionados ao DF5_docentes_BR os campos CD_AREA_BASICA_TITULACAO_NIVEL_1 e NM_AREA_BASICA_TITULACAO_NIVEL_1, registrando a Grande Área de titulação de cada orientador. Essa organização permite futuras análises comparativas entre a formação dos

orientadores e a área de enquadramento dos programas nos quais atuam, além de facilitar a geração de representações gráficas. O processo de inserção e ajuste desses dados no DF5_docentes_BR foi implementado por meio do *script* descrito no Quadro 79.

Quadro 79 – Script desenvolvido para inserir dois campos adicionais ao DF5_docentes_BR

```
import pandas as pd
# Caminho do arquivo
file_path = '/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF5/DF5.csv'
# Ler o arquivo
df = pd.read_csv(file_path, encoding="utf-8", sep=";")
# Identificar o índice da coluna 'NM AREA BASICA TITULACAO'
nm_index = df.columns.get_loc("NM AREA BASICA TITULACAO") + 1
# Adicionar novas colunas após 'NM AREA BASICA TITULACAO'
df.insert(nm_index, 'CD AREA BASICA TITULACAO NIVEL_1', pd.NA) # Insere coluna com valores NA
df.insert(nm_index + 1, 'NM AREA BASICA TITULACAO NIVEL_1', pd.NA) # Insere coluna com valores NA
# Salvar o DataFrame atualizado no mesmo arquivo
df.to_csv(file_path, index=False, encoding='utf-8', sep=';')
print("Novas colunas adicionadas e arquivo salvo com sucesso.")
```

Fonte: Dados de pesquisa (2024).

Após a criação do arquivo CSV pelo *script* do Quadro 79, os dados foram organizados em ordem alfabética para facilitar a análise. Cada nome de área básica foi verificado e comparado com a documentação oficial da Capes, para garantir a correta associação ao código correspondente e à Grande Área.

Com base nessa verificação, foi desenvolvido um novo *script*, apresentado no Quadro 80, que processa o DataFrame utilizando um dicionário de atualizações predefinido, no qual cada chave representa um valor do campo NM_AREA_BASICA_TITULACAO, e os valores contêm as modificações e dados corrigidos. O DataFrame é carregado, atualizado e salvo novamente no arquivo CSV, garantindo a consistência dos dados no DF5_docentes_BR conforme as diretrizes da Capes.

Quadro 80 – Script desenvolvido para inserir valores aos campos de código e nome de titulação de nível 1

```
import pandas as pd
def update_fields(df, updates):
    for search_value, change_details in updates.items():
        # Encontrar as linhas que correspondem ao valor de busca
        mask = df['NM AREA BASICA TITULACAO'] == search_value

        # Aplicar as atualizações para as linhas correspondentes
        for field, new_value in change_details.items():
            df.loc[mask, field] = new_value
    return df
# Caminho do arquivo
file_path = '/content/drive/MyDrive/CAPES/datasets/DF5/DF5.csv'
# Ler o arquivo
df = pd.read_csv(file_path, encoding="utf-8", sep=";")
# Dicionário com as atualizações a serem feitas
updates = {
    "ACÚSTICA": {
        'CD AREA BASICA TITULACAO': 10502033,
        'CD AREA BASICA TITULACAO NIVEL_1': 10000003,
```

```

'NM_AREA_BASICA_TITULACAO_NIVEL_1': 'CIÊNCIAS EXATAS E DA TERRA'
},
"ADEQUAÇÃO AMBIENTAL": {
'CD_AREA_BASICA_TITULACAO': 60403012,
'CD_AREA_BASICA_TITULACAO_NIVEL_1': 60000007,
'NM_AREA_BASICA_TITULACAO_NIVEL_1': 'CIÊNCIAS SOCIAIS APLICADAS'
}
# Adicione mais atualizações conforme necessário
}
# Atualizar o DataFrame
df = update_fields(df, updates)
# Salvar o DataFrame atualizado
df.to_csv(file_path, index=False, encoding='utf-8', sep=';')
print("Atualizações aplicadas e arquivo salvo com sucesso.")

```

Fonte: Dados de pesquisa (2024).

Foram realizadas intervenções nos dados gerados pelo *script* do Quadro 80, com verificações no currículo Lattes para atualizar as áreas do conhecimento de docentes com registros incompletos. No entanto, algumas limitações foram identificadas. O preenchimento da área de formação no Lattes não é obrigatório, resultando em registros ausentes. Além disso, nomes que não correspondiam a docentes brasileiros geraram múltiplas ocorrências ou associações a indivíduos de outras áreas. Para casos em que a área de conhecimento não pôde ser determinada, foi atribuído o código '99999999' para a área e a subárea, com a descrição 'não informada'. Essa padronização mantém a consistência dos dados no DF5_docentes_BR e facilita análises futuras. A Tabela 25 apresenta os dados ajustados finais do DF5_docentes_BR.

Tabela 25 – Dados finais do DF5_docentes_BR após os ajustes realizados

Índice	Campo	Antes	Depois	Valores únicos antes	Valores únicos depois
1	ANO_BASE	1482960	1482960	19	19
2	CD_PROGRAMA	1482960	1482960	5151	5151
3	NM_PROGRAMA	1482960	1482960	2554	2554
4	ID_DOCENTE	991187	1482960	109877	123618
5	NM_DOCENTE	1482960	1482960	145196	145080
6	NM_NACIONALIDADE_DOCENTE	1482960	1482960	156	129
7	AN_NASCIMENTO_DOCENTE	1482952	1482960	89	89
8	DS_CATEGORIA_DOCENTE	1482960	1482960	3	3
9	DS_TIPO_VINCULO_DOCENTE	1482941	1482960	7	7
10	NM_ULTIMO_GRAU_TITULACAO	1482946	1482960	7	7
11	AN_ULTIMO_GRAU_TITULACAO	1482946	1482960	83	83

Índice	Campo	Antes	Depois	Valores únicos antes	Valores únicos depois
12	SG_IES_ULTIMO_GRAU_TITULACAO	1469108	1479982	7011	1980
13	NM_IES_GRAU_TITULACAO	1482946	1482960	14748	2162
14	NM_PAIS_ULTIMO_GRAU_TITULACAO	1482946	1482960	109	82
15	CD_AREA_BASICA_TITULACAO	991173	1482960	2495	1156
16	NM_AREA_BASICA_TITULACAO	1482946	1482960	1197	1156
17	CD_AREA_BASICA_TITULACAO_NIVEL_1	0	1482960	0	10
18	NM_AREA_BASICA_TITULACAO_NIVEL_1	0	1482960	0	10
19	NM_DOCENTE_S_ACENTO	0	1482960	0	123319

Fonte: Dados de pesquisa (2024).

As atualizações apresentadas na Tabela 25 incluíram ajustes e padronizações para adequar os registros às necessidades da pesquisa. Alguns campos passaram por intervenções: no campo ID_DOCENTE, foram inseridos códigos para docentes sem identificação, resultando na fusão e alteração de alguns registros. O campo CD_AREA_BASICA_TITULACAO recebeu códigos adicionais para corrigir lacunas. No campo NM_DOCENTE, foram aplicadas padronizações nos nomes, com a criação do campo NM_DOCENTE_S_ACENTO para facilitar a verificação de acentuação. Já no campo NM_NACIONALIDADE_DOCENTE, diferentes grafias de um mesmo país foram unificadas. Essas modificações aprimoraram a estrutura do DF5_docentes_BR, permitindo refinamentos no DF3_trabalhos_BR.

Os ajustes realizados no DF5_docentes_BR viabilizaram a padronização e organização dos dados, facilitando uma análise mais precisa dos docentes nos programas de pós-graduação *stricto sensu* no Brasil. Além disso, reforçaram a importância da gestão consistente de dados na pesquisa acadêmica. Esta seção documenta o processo de manipulação dos registros e demonstra como a padronização e a verificação detalhada contribuem para a qualidade e confiabilidade das informações utilizadas nos estudos.

Finalizado o trabalho com os registros dos docentes, passa-se agora à análise dos conjuntos de dados dos discentes.

Passo 3: Entendimento e ajustes do conjunto de dados dos Discentes

Este passo 3, envolve a análise dos dados dos discentes dos Programas de Pós-Graduação *stricto sensu* no Brasil. Esse conjunto compreende 19 arquivos, de 2004 a 2022, contendo informações sobre nome, nacionalidade, situação acadêmica, nível do curso e orientador principal, conforme informado pelos PPG. Inicialmente, é realizada uma análise para compreender a estrutura de cada arquivo disponibilizado pela CAPES, que permite organizar o DF6_discentes_BR, garantindo a padronização e integridade dos registros para futuras análises. A seguir apresentam-se o entendimento (item 1) e os ajustes (item 2) do conjunto de dados dos discentes.

Entendimento do conjunto de dados dos Discentes

Na análise dos dados dos discentes, os 19 arquivos foram abertos no LibreOffice Calc para visualização. Esses arquivos estão divididos em dois períodos: de 2004 a 2012, com 31 colunas, e de 2013 a 2022, com 37 colunas. A Tabela 26 apresenta o volume total de dados por campo em ambos os períodos, permitindo avaliar as variáveis que mantêm correspondência ao longo do tempo.

Tabela 26 – Total dos dados dos conjuntos de dados referentes aos discentes

2004-2012		2013-2022	
Campo	Linhas	Campo	Linhas
AN_BASE	2480971	AN_BASE	3731898
AN_MATRICULA_DISCENTE	2480971	AN_NASCIMENTO_DISCENTE	3731898
AN_NASCIMENTO_DISCENTE	2480971	CD_AREA_AVALIACAO	3731898
AN_SITUACAO_DISCENTE	2480971	CD_CONCEITO_CURSO	3731898
CD_AREA_AVALIACAO	2480971	CD_CONCEITO_PROGRAMA	3731898
CD_PROGRAMA_IES	2480971	CD_ENTIDADE_CAPES	3731898
CS_STATUS_JURIDICO	2480971	CD_ENTIDADE_EMEC	3489720
DS_FAIXA_ETARIA_DISCENTE	2476031	CD_PROGRAMA_IES	3731898
DT_TESE DISSERTACAO	423505	CS_STATUS_JURIDICO	3731898
ME_MATRICULA_DISCENTE	2480971	DS_DEPENDENCIA_ADMINISTRATIVA	3731898
ME_SITUACAO_DISCENTE	2480971	DS_FAIXA_ETARIA	3731898
NM_AREA_AVALIACAO	2480971	DS_GRAU_ACADEMICO_DISCENTE	3731898
NM_DISCENTE	2480971	DS_TIPO_NACIONALIDADE_DISCENTE	3731898
NM_ENTIDADE_ENSINO	2480971	DT_MATRICULA_DISCENTE	3731898
NM_MODALIDADE_PROGRAMA	2480971	DT_SITUACAO_DISCENTE	3731898
NM_NIVEL_CONCLUSAO_DISCENTE	2480971	ID_ADD_FOTO_PROGRAMA	3731898

2004-2012		2013-2022	
Campo	Linhas	Campo	Linhas
NM_NIVEL_PROGRAMA	1849570	ID_ADD_FOTO_PROGRAMA_IES	3731898
NM_NIVEL_TITULACAO_DISCENTE	2480971	ID_PESSOA	3731898
NM_ORIENTADOR_PRINCIPAL	2480971	NM_AREA_AVALIACAO	3731898
NM_PAIS_ORIGEM_DISCENTE	2480971	NM_DISCENTE	3731898
NM_PROGRAMA_IES	2480971	NM_ENTIDADE_ENSINO	3731898
NM_REGIAO_ENTIDADE	2480971	NM_GRANDE_AREA_CONHECIMENTO	3731898
NM_SITUACAO_DISCENTE	2480971	NM_GRAU_PROGRAMA	3731898
NM_TESE DISSERTACAO	423505	NM_MODALIDADE_PROGRAMA	3731898
NM_TIPO_DOCENTE_ORIENT_PRINC	2480971	NM_MUNICIPIO_PROGRAMA_IES	3731898
NR_IDADE_DISCENTE	2476034	NM_ORIENTADOR	1313444
NR_SEQ_ORIENTADOR_PRINCIPAL	2480971	NM_ORIENTADOR_PRINCIPAL	2418454
NR_SEQUENCIAL_DISCENTE	2480971	NM_PAIS_NACIONALIDADE_DISCENTE	3731898
NR_SEQUENCIAL_TESE	2480971	NM_PROGRAMA_IES	3731898
SG_ENTIDADE_ENSINO	2480971	NM_REGIAO	3731898
SG_UF_ENTIDADE_ENSINO	2480971	NM_SITUACAO_DISCENTE	3731898
		NM_TESE DISSERTACAO	3731898
		NR_DOCUMENTO_DISCENTE	3731898
		QT_MES_TITULACAO	3731898
		SG_ENTIDADE_ENSINO	3731898
		SG_UF_PROGRAMA	3731898
		ST_INGRESSANTE	3731898
		TP_DOCUMENTO_DISCENTE	3731898

Fonte: Dados de pesquisa (2024).

O objetivo dessa análise é definir a melhor forma de estruturar e limpar os dados, garantindo a consolidação das informações dos discentes de maneira completa e consistente.

A análise dos dados da Tabela 26 identificou atributos comuns entre os dois conjuntos de dados (2004 a 2012 e 2013 a 2022), incluindo ano base, data de matrícula (mês e ano no primeiro conjunto), ano de nascimento, código da área de avaliação, código do programa, status jurídico do discente, faixa etária, data de finalização do trabalho, nome do discente, instituição de ensino, modalidade e nível do programa, nome do orientador, país de origem do discente, além da localização do programa (região, estado, nome e sigla da IES). Porém, os dados do primeiro conjunto estão estruturados de maneira diferente daqueles do segundo

conjunto e, além disso, o primeiro conjunto possui um número menor de colunas. Para resolver esse problema, foi necessário estabelecer uma correlação direta entre os dois conjuntos. A Tabela 27 apresenta o resultado desse alinhamento, apresentando as 13 colunas que irão compor o DF6_discentes_BR.

Tabela 27 – Campos que vão constituir o DF6_discentes_BR

Campo
AN_BASE
AN_NASCIMENTO_DISCENTE
CD_PROGRAMA_IES
ID_PESSOA
NM_DISCENTE
DT_MATRICULA_DISCENTE
DT_SITUACAO_DISCENTE
NM_ENTIDADE_ENSINO
NM_PAIS_NACIONALIDADE_DISCENTE
NM_ORIENTADOR
DS_FAIXA_ETARIA
DS_GRAU_ACADEMICO_DISCENTE
QT_MES_TITULACAO

Fonte: Dados de pesquisa (2024).

Essa estrutura para o DF6_discentes_BR busca garantir a padronização e a consistência dos registros nos dois conjuntos de dados, facilitando análises comparativas e estatísticas sobre os discentes nos programas de pós-graduação brasileira.

O Quadro 81 apresenta o *script* desenvolvido para a criação do DF6_discentes_BR, estabelecendo a estrutura inicial do DataFrame que armazenará os dados dos discentes dos programas de pós-graduação *stricto sensu*.

Quadro 81 – Script desenvolvido para criação do DF6_discentes_BR

```
import pandas as pd
# Criar um DataFrame vazio com as colunas especificadas
colunas = [
    "AN_BASE", "AN_NASCIMENTO_DISCENTE", "CD_PROGRAMA_IES", "ID_PESSOA", "NM_DISCENTE", "DT_MATRICULA_DISCENTE", "DT_SITUACAO_DISCENTE", "NM_ENTIDADE_ENSINO", "NM_PAIS_NACIONALIDADE_DISCENTE", "NM_ORIENTADOR", "DS_FAIXA_ETARIA", "DS_GRAU_ACADEMICO_DISCENTE", "QT_MES_TITULACAO", "NM_SITUACAO_DISCENTE"
]
df6 = pd.DataFrame(columns=colunas)
```

```
# Salvar o DataFrame como um arquivo CSV
df6.to_csv("/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF6/DF6.csv", index=False, encoding="utf-8", sep=";")
```

Fonte: Dados de pesquisa (2024).

Com a estrutura do DF6_discentes_BR definida, foi necessário tratar inconsistências nos dados antes da integração final, especificamente nos campos DT_MATRICULA_DISCENTE, DT_SITUACAO_DISCENTE e QT_MES_TITULACAO. Nos dois primeiros campos, os dados estavam armazenados de forma heterogênea nos arquivos originais, exigindo um ajuste para padronização. Para isso, foi desenvolvido um script, apresentado no Quadro 82, que atribui o dia "1" a todas as datas, garantindo um formato uniforme e permitindo cálculos posteriores. Para o terceiro campo (QT_MES_TITULACAO), uma das principais operações realizadas pelo *script* é a determinação do tempo até a titulação, o que não estava disponível no primeiro conjunto de dados (2004-2012). Para esses registros, a quantidade de meses até a conclusão é calculada com base na diferença entre a data de matrícula e a de conclusão.

Quadro 82 – Script desenvolvido para inserção dos dados de 2004 a 2012

```
import pandas as pd
# Defina os caminhos dos arquivos
source_file_path = '/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF6/2004.csv'
destination_file_path = '/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF6/DF6.csv'
# Leia o arquivo de origem
df = pd.read_csv(source_file_path, sep=';', encoding='ISO-8859-1')
# Filtre os dados conforme o valor de NM_SITUACAO_DISCENTE
filtered_df = df[df['NM_SITUACAO_DISCENTE'].isin(['TITULADO', 'MUDANÇA DE NÍVEL COM DEFESA', 'MUDANÇA DE NÍVEL SEM DEFESA'])]
# Renomeie as colunas conforme especificado
filtered_df = filtered_df.rename(columns={
    'AN_BASE': 'AN_BASE', 'AN_NASCIMENTO_DISCENTE': 'AN_NASCIMENTO_DISCENTE', 'CD_PROGRAMA_IES':
    'CD_PROGRAMA_IES', 'ID_PESSOA': 'NR_SEQUENCIAL_DISCENTE',
    'NM_DISCENTE': 'NM_DISCENTE', 'NM_ENTIDADE_ENSINO': 'NM_ENTIDADE_ENSINO',
    'NM_PAIS_NACIONALIDADE_DISCENTE': 'NM_PAIS_ORIGEM_DISCENTE', 'NM_ORIENTADOR':
    'NM_ORIENTADOR_PRINCIPAL', 'DS_FAIXA_ETARIA': 'DS_FAIXA_ETARIA_DISCENTE',
    'DS_GRAU_ACADEMICO_DISCENTE': 'NM_NIVEL_CONCLUSAO_DISCENTE'
})
# Função para criar datas a partir de ano e mês, lidando com valores inválidos
def create_date(year, month):
    try:
        year = int(year)
        month = int(month)
        if year > 1900 and year < 2100 and 1 <= month <= 12:
            return pd.Timestamp(year=year, month=month, day=1)
    except:
        return pd.NaT
    return pd.NaT
# Função para calcular faixa etária
def calcular_faixa_etaria(an_base, an_nascimento):
    try:
        idade = an_base - an_nascimento
        if idade <= 19:
            return '19 OU MENOS'
        elif 20 <= idade <= 24:
            return '20 A 24 ANOS'
        elif 25 <= idade <= 29:
            return '25 A 29 ANOS'
        elif 30 <= idade <= 34:
```

```

    return '30 A 34 ANOS'
elif 35 <= idade <= 39:
    return '35 A 39 ANOS'
elif 40 <= idade <= 44:
    return '40 A 44 ANOS'
elif 45 <= idade <= 49:
    return '45 A 49 ANOS'
elif 50 <= idade <= 54:
    return '50 A 54 ANOS'
elif 55 <= idade <= 59:
    return '55 A 59 ANOS'
elif 60 <= idade <= 64:
    return '60 A 64 ANOS'
elif 65 <= idade <= 69:
    return '65 A 69 ANOS'
elif idade >= 70:
    return '70 OU MAIS'
except:
    return pd.NA
return pd.NA
# Crie as colunas de datas a partir dos campos AN_MATRICULA_DISCENTE e ME_MATRICULA_DISCENTE
filtered_df['DT_MATRICULA_DISCENTE'] = filtered_df.apply(lambda row: create_date(row['AN_MATRICULA_DISCENTE'],
row['ME_MATRICULA_DISCENTE']), axis=1)
# Crie as colunas de datas a partir dos campos AN_SITUACAO_DISCENTE e ME_SITUACAO_DISCENTE
filtered_df['DT_SITUACAO_DISCENTE'] = filtered_df.apply(lambda row: create_date(row['AN_SITUACAO_DISCENTE'],
row['ME_SITUACAO_DISCENTE']), axis=1)
# Preencha o campo DS_FAIXA_ETARIA_DISCENTE se estiver vazio
filtered_df['DS_FAIXA_ETARIA_DISCENTE'] = filtered_df.apply(
    lambda row: row['DS_FAIXA_ETARIA_DISCENTE'] if pd.notna(row['DS_FAIXA_ETARIA_DISCENTE']) else
    calcular_faixa_etaria(row['AN_BASE'], row['AN_NASCIMENTO_DISCENTE']),
    axis=1
)
# Remova linhas com datas inválidas
filtered_df = filtered_df.dropna(subset=['DT_MATRICULA_DISCENTE', 'DT_SITUACAO_DISCENTE'])
# Calcule a diferença em meses entre DT_SITUACAO_DISCENTE e DT_MATRICULA_DISCENTE
filtered_df['QT_MES_TITULACAO'] = ((filtered_df['DT_SITUACAO_DISCENTE'] -
filtered_df['DT_MATRICULA_DISCENTE']).dt.days // 30).astype('Int64')
# Selecione e ordene as colunas conforme necessário
filtered_df = filtered_df[[
    'AN_BASE', 'AN_NASCIMENTO_DISCENTE', 'CD_PROGRAMA_IES', 'NR_SEQUENCIAL_DISCENTE', 'NM_DISCENTE',
    'DT_MATRICULA_DISCENTE',
    'DT_SITUACAO_DISCENTE', 'NM_ENTIDADE_ENSINO', 'NM_PAIS_ORIGEM_DISCENTE',
    'NM_ORIENTADOR_PRINCIPAL',
    'DS_FAIXA_ETARIA_DISCENTE', 'NM_NIVEL_CONCLUSAO_DISCENTE', 'QT_MES_TITULACAO'
]]
# Salve o DataFrame resultante no arquivo de destino
filtered_df.to_csv(destination_file_path, sep=';', encoding='utf-8', index=False)
print("Arquivo salvo com sucesso!")

```

Fonte: Dados de pesquisa (2024).

Além da padronização de datas, outros campos, como nome e sigla da IES, região, estado e modalidade do programa, já estavam disponíveis em outros DataFrames. Para evitar redundância e garantir a integridade das informações, esses dados foram cruzados e unificados ao DF6_discentes_BR.

Outro desafio identificado foi o campo DS_FAIXA_ETARIA_DISCENTE, que apresentou lacunas em alguns registros. Para corrigir essa inconsistência, o script implementado no Quadro 81 incluiu uma função que calcula e insere automaticamente a faixa etária dos discentes, garantindo a completude dos dados e a consistência do DF6_discentes_BR.

Além dos ajustes estruturais, o script também definiu critérios para a seleção das linhas a serem incorporadas ao DF6_discentes_BR. No campo NM_SITUACAO_DISCENTE, os registros incluíam diferentes *status*, como MATRICULADO, DESLIGADO, ABANDONO, TITULADO, MUDANÇA DE NÍVEL COM DEFESA e MUDANÇA DE NÍVEL SEM DEFESA. Para este estudo, foram considerados apenas os discentes com *status* TITULADO, MUDANÇA DE NÍVEL COM DEFESA e MUDANÇA DE NÍVEL SEM DEFESA. Essa seleção prioriza informações relevantes para a análise da trajetória acadêmica, focando nos dados que permitem avaliar a progressão dos discentes nos programas.

Devido ao grande volume de arquivos, optou-se por um processamento segmentado, lendo e gravando cada arquivo individualmente. Essa abordagem facilita a validação dos dados e evita dificuldades na manipulação de grandes volumes de dados diretamente em planilhas. Também garante maior precisão nas verificações e consolidação dos registros no DF6_discentes_BR.

A Tabela 28 apresenta os dados consolidados referentes ao período de 2004 a 2012 do DF6_discentes_BR, incluindo a quantidade de registros por campo e os valores únicos encontrados, permitindo uma visão geral da estrutura e variabilidade das informações.

Tabela 28 – Resultado do DF6_discentes_BR com os dados de 2004 a 2012

Campo	Quantidade de linhas	Valores únicos
AN_BASE	425929	9
AN_NASCIMENTO_DISCENTE	425929	94
CD_PROGRAMA_IES	425929	3102
ID_PESSOA	425929	14791
NM_DISCENTE	425929	389826
DT_MATRICULA_DISCENTE	425929	217
DT_SITUACAO_DISCENTE	425929	168
NM_ENTIDADE_ENSINO	425929	351
NM_PAIS_NACIONALIDADE_DISCENTE	425929	126
NM_ORIENTADOR	425929	56773
DS_FAIXA_ETARIA	425929	12
DS_GRAU_ACADEMICO_DISCENTE	425929	4
QT_MES_TITULACAO	425929	151

Fonte: Dados de pesquisa (2024).

Numa visão geral, a Tabela 28 mostra que o DF6_discentes_BR contém 425.929 registros, abrangendo nove anos de dados (2004-2012). O campo NM_DISCENTE apresenta quase 390 mil valores únicos, indicando um alto nível de individualização entre os discentes. No entanto, o campo ID_PESSOA, que deveria servir como identificador único, contém apenas 14.791 valores distintos, sugerindo a possibilidade de registros duplicados, inconsistências nos identificadores ou múltiplas matrículas associadas ao mesmo discente. A presença de 3.102 códigos únicos no campo CD_PROGRAMA_IES reflete a diversidade dos programas de pós-graduação. O número de 94 valores únicos em AN_NASCIMENTO_DISCENTE é compatível com o esperado, considerando diferentes idades de ingresso na pós-graduação ao longo dos anos. Esses resultados destacam a necessidade de verificar possíveis inconsistências nos identificadores e avaliar a distribuição dos dados em relação às trajetórias acadêmicas dos discentes.

Em seguida, inicia-se o processamento do segundo conjunto de dados (2013-2022). O *script* do Quadro 83 extrai e integra essas informações ao DF6_discentes_BR, utilizando o campo NM_SITUACAO_DISCENTE como critério de filtragem, conforme aplicado no primeiro conjunto. Além disso, foi necessário padronizar as datas, convertendo formatos inconsistentes, como '21FEV1300:00:00' para '2013-02-21', garantindo uniformidade na estrutura dos dados.

Quadro 83 – Script desenvolvido para inserção dos dados de 2013 a 2022

```
import pandas as pd
from datetime import datetime
import os
# Função para converter a data do formato DDMMAA para YYYY-MM-DD
def convert_date(date_str):
    try:
        # Apenas pegue os primeiros 7 caracteres para ignorar a parte de tempo
        date_str = date_str[:7]
        return datetime.strptime(date_str, '%d%b%y').strftime('%Y-%m-%d')
    except ValueError:
        return pd.NaT
# Caminhos dos arquivos
folder_path = '/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF6'
destination_file_path = os.path.join(folder_path, 'DF6.csv')
# Mapeamento de colunas
columns_mapping = {'AN_BASE': 'AN_BASE', 'AN_NASCIMENTO_DISCENTE': 'AN_NASCIMENTO_DISCENTE',
'CD_PROGRAMA_IES': 'CD_PROGRAMA_IES', 'ID_PESSOA': 'ID_PESSOA', 'NM_DISCENTE': 'NM_DISCENTE',
'DT_MATRICULA_DISCENTE': 'DT_MATRICULA_DISCENTE', 'DT_SITUACAO_DISCENTE': 'DT_SITUACAO_DISCENTE',
'NM_ENTIDADE_ENSINO': 'NM_ENTIDADE_ENSINO', 'NM_PAIS_NACIONALIDADE_DISCENTE':
'NM_PAIS_ORIGEM_DISCENTE', 'NM_ORIENTADOR': 'NM_ORIENTADOR', 'DS_FAIXA_ETARIA':
'DS_FAIXA_ETARIA', 'DS_GRAU_ACADEMICO_DISCENTE':
'NM_NIVEL_CONCLUSAO_DISCENTE', 'QT_MES_TITULACAO': 'QT_MES_TITULACAO'
}
# Inicialize um DataFrame vazio para armazenar todos os dados
df_all_years = pd.DataFrame()
# Loop para processar arquivos de 2013 a 2022
for year in range(2013, 2023):
    file_path = os.path.join(folder_path, f'{year}.csv')
    # Leia o arquivo de origem com dtype explicitamente especificado para evitar o aviso de tipos mistos
    df = pd.read_csv(file_path, sep=';', encoding='ISO-8859-1', dtype=str)
```

```

# Filtre os dados conforme o valor de NM_SITUACAO_DISCENTE
filtered_df = df[df['NM_SITUACAO_DISCENTE'].isin(['TITULADO', 'MUDANÇA DE NÍVEL COM DEFESA', 'MUDANÇA DE NÍVEL SEM DEFESA'])]
# Ajustar o mapeamento para a coluna do orientador de acordo com o ano
if year <= 2016:
    filtered_df = filtered_df.rename(columns={'NM_ORIENTADOR': 'NM_ORIENTADOR'})
else:
    filtered_df = filtered_df.rename(columns={'NM_ORIENTADOR_PRINCIPAL': 'NM_ORIENTADOR'})
# Verifique a presença de cada coluna no DataFrame antes de renomear e selecionar
available_columns = {col: columns_mapping[col] for col in columns_mapping if col in filtered_df.columns}
available_columns['NM_ORIENTADOR'] = 'NM_ORIENTADOR'
# Renomeie as colunas conforme o mapeamento disponível
filtered_df = filtered_df.rename(columns=available_columns)
# Selecione apenas as colunas mapeadas disponíveis
filtered_df = filtered_df[list(available_columns.values())]
# Converta as colunas de datas para o formato YYYY-MM-DD
if 'DT_MATRICULA_DISCENTE' in filtered_df.columns:
    filtered_df['DT_MATRICULA_DISCENTE'] = filtered_df['DT_MATRICULA_DISCENTE'].apply(convert_date)
if 'DT_SITUACAO_DISCENTE' in filtered_df.columns:
    filtered_df['DT_SITUACAO_DISCENTE'] = filtered_df['DT_SITUACAO_DISCENTE'].apply(convert_date)
# Adicione os dados processados ao DataFrame principal
df_all_years = pd.concat([df_all_years, filtered_df])
# Validação de campos
print(f"Validação do arquivo {year}.csv:")
for column in filtered_df.columns:
    num_nulls = filtered_df[column].isnull().sum()
    num_zeros = (filtered_df[column] == '0').sum() if filtered_df[column].dtype == 'object' else (filtered_df[column] == 0).sum()
    num_empty = (filtered_df[column] == "").sum()
    print(f"Coluna '{column}': {num_nulls} nulos, {num_empty} vazios, {num_zeros} zerados")
# Verifique se o arquivo de destino já existe
if os.path.exists(destination_file_path):
    # Se existir, carregue os dados existentes
    df_existing = pd.read_csv(destination_file_path, sep=';', encoding='utf-8')
    # Concatene os novos dados com os existentes
    df_all_years = pd.concat([df_existing, df_all_years])
# Verifique quantos campos estão sendo levados com valores nulos, vazios ou zerados no DataFrame final
print("Validação do DataFrame final:")
for column in df_all_years.columns:
    num_nulls = df_all_years[column].isnull().sum()
    num_zeros = (df_all_years[column] == '0').sum() if df_all_years[column].dtype == 'object' else (df_all_years[column] == 0).sum()
    num_empty = (df_all_years[column] == "").sum()
    print(f"Coluna '{column}': {num_nulls} nulos, {num_empty} vazios, {num_zeros} zerados")
# Salve o DataFrame resultante no arquivo de destino
df_all_years.to_csv(destination_file_path, sep=';', encoding='utf-8', index=False)
print("Arquivo salvo com sucesso!")

```

Fonte: Dados de pesquisa (2024).

Durante a execução do *script* do Quadro 83, foram identificadas variações na nomenclatura das colunas ao longo dos anos, exigindo ajustes no processamento. O campo do orientador, por exemplo, foi registrado como 'NM_ORIENTADOR' de 2013 a 2016 e 'NM_ORIENTADOR_PRINCIPAL' de 2017 a 2022. Essa diferença só foi detectada após a integração inicial, demandando modificações no script para garantir a correta correspondência dos dados. A versão revisada do script assegura a recuperação completa dos registros e sua inserção correta no DF6_discentes_BR. A Tabela 29 apresenta a consolidação dos dados, incluindo a contagem total de registros por campo e os valores únicos identificados.

Tabela 29 – Resultado do DF6_discentes_BR com os dados de 2004 a 2022

Campo	Quantidade de linhas	Valores únicos
AN_BASE	812046	10

Campo	Quantidade de linhas	Valores únicos
AN_NASCIMENTO_DISCENTE	812046	70
CD_PROGRAMA_IES	812046	4803
ID_PESSOA	812046	724003
NM_DISCENTE	812046	718542
DT_MATRICULA_DISCENTE	812046	2870
DT_SITUACAO_DISCENTE	812046	1781
NM_ENTIDADE_ENSINO	812046	642
NM_PAIS_NACIONALIDADE_DISCENTE	812046	126
NM_ORIENTADOR	812046	90894
DS_FAIXA_ETARIA	812046	12
DS_GRAU_ACADEMICO_DISCENTE	812046	4
QT_MES_TITULACAO	812046	160

Fonte: Dados de pesquisa (2024).

A Tabela 29 apresenta a distribuição e a variedade das informações no DF6_discentes_BR, servindo de base para a análise dos discentes em programas de pós-graduação *stricto sensu*. A próxima ação é revisar cada coluna, aplicando ajustes para garantir a completude e a precisão dos dados. Esse refinamento ajuda na integração do DF6_discentes_BR com outros conjuntos de dados do estudo e para sua utilização em pesquisas futuras, assegurando a qualidade e a aplicabilidade do DataFrame na análise dos perfis e trajetórias dos discentes.

Tendo finalizado o entendimento dos dados sobre os discentes, passa-se para os ajustes mais refinados.

Ajustes do conjunto de dados dos Discentes

Nos ajustes mais refinados, cada campo é tratado individualmente. Inicia-se pelo campo AN_BASE, que contém 19 valores únicos, correspondentes aos anos de 2004 a 2022, conforme apresentado na Tabela 30.

Tabela 30 – Dados do campo AN_BASE

AN_BASE	Quantidade
2004	35092
2005	40041

AN_BASE	Quantidade
2006	41926
2007	43092
2008	47062
2009	50465
2010	51132
2011	55807
2012	61312
2013	67760
2014	70929
2015	76129
2016	80555
2017	86391
2018	91765
2019	95753
2020	80114
2021	79985
2022	82665

Fonte: Dados de pesquisa (2024).

O conjunto de dados do campo AN_BASE possui 1.237.975 registros no período analisado. A distribuição temporal mostra um aumento no número de registros de 2004 a 2019, seguido por uma queda em 2020 e 2021, com recuperação em 2022. Essa redução observada nos anos de pandemia de COVID-19 reflete os impactos desse período nas atividades acadêmicas, possivelmente devido a questões como restrições institucionais, atraso nas defesas de dissertações e teses, mudanças nas dinâmicas de pesquisa, dificuldades operacionais e menor inserção de novos discentes enfrentadas pelos programas de pós-graduação. Verificou-se que o campo AN_BASE não apresenta discrepâncias que comprometam a validade dos dados, garantindo sua consistência para análises longitudinais.

A análise prossegue para o campo AN_NASCIMENTO_DISCENTE, quando foram identificados 98 valores únicos, todos de quatro dígitos. No entanto, quatro registros apresentam o valor '0', sugerindo erro de entrada. Além disso, anos entre 1900 e 1920 possuem no máximo três registros anuais, enquanto os anos de 2000 a 2016 apresentam até dois registros por ano. A validação desses dados é limitada devido à proteção do ano de

nascimento pela LGPD (Rapôso, 2019), impossibilitando a verificação direta. Assim, os dados são mantidos como estão, e futuras análises ou representações gráficas deverão incluir notas explicativas sobre possíveis imprecisões, especialmente nos casos com valor '0', nas datas entre 1900 e 1920 e nos anos posteriores a 2000, que somam 28 registros (0,0022% do total).

A análise segue para o campo `CD_PROGRAMA_IES`, que registra o código do programa na instituição. O campo contém 5.047 valores únicos distribuídos em 1.237.975 linhas, sem valores nulos ou fora do padrão. O formato é alfanumérico, incluindo códigos com letras, como "20001010053P3", não são necessárias alterações neste campo.

Logo após, segue-se para o campo `ID_PESSOA`, que identifica os discentes nos registros da Capes. De 2004 a 2012, essa identificação não era aplicada, e cada discente recebia um número sequencial em cada programa. O campo possui 1.237.975 entradas e 738.705 valores únicos, mas a repetição de IDs sugere múltiplos registros para um mesmo discente. Embora se espere um ID por mestrado e outro por doutorado, alguns discentes podem possuir múltiplos títulos nesses níveis. Esse campo é utilizado para junções com outros DataFrames e deve ser tratado com cautela para garantir a precisão das análises.

A análise segue para o campo `NM_DISCENTE`, que armazena os nomes dos discentes, que contém 1.237.975 registros e 1.032.744 valores únicos. A presença de homônimos e variações no uso de acentos e caracteres especiais afeta a contagem de valores únicos e dificulta a identificação em buscas e no currículo Lattes. Para resolver essas questões, foi criado um script, ilustrado no Quadro 81, que usa a função *unidecode* para remover acentos e caracteres especiais, resultando em 1.006.849 valores únicos. Além disso, uma análise revelou prefixos como (H), (E) ou (G) antes de alguns nomes, que foram removidos, ajustando o total de valores únicos para 1.006.847.

O campo `NM_DISCENTE`, que armazena os nomes dos discentes, contém 1.237.975 registros e 1.032.744 valores únicos. A presença de homônimos e variações no uso de acentos e caracteres especiais afeta a contagem de valores únicos e dificulta a identificação em buscas e no currículo Lattes. Para padronizar os registros, foi desenvolvido um *script*, descrito no Quadro 84, que utiliza a função *unidecode* para remover acentos e caracteres especiais, reduzindo o total para 1.006.849 valores únicos.

Quadro 84 – Script para criação do campo `NM_DISCENTE_SEM_ACENTO` no `DF6_discentes_BR` e migração dos valores do `NM_DISCENTE`

```
!pip install unidecode
import pandas as pd
```

```

from unidecode import unidecode
# Caminho do arquivo
file_path = '/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF6/DF6.csv'
# Leia o arquivo CSV
df = pd.read_csv(file_path, sep=';', encoding='utf-8')
# Crie a nova coluna NM_DISCENTE_SEM_ACENTO removendo acentos e caracteres especiais
df['NM_DISCENTE_SEM_ACENTO'] = df['NM_DISCENTE'].apply(unidecode)
# Salve o DataFrame atualizado de volta ao arquivo CSV
df.to_csv(file_path, sep=';', encoding='utf-8', index=False)
print("Arquivo atualizado com sucesso!")

```

Fonte: Dados de pesquisa (2024).

A análise com esse *script* identificou e removeu prefixos como (H), (E) e (G) antes de alguns nomes, ajustando o total de valores únicos para 1.006.847. Com esses ajustes, o campo NM_DISCENTE foi refinado, tornando-se mais adequado para análises futuras, com uma base de dados mais organizada.

A análise do DF6_discentes_BR dedicou atenção aos campos data de matrícula e data da situação, que indicam a data de titulação dos discentes. Optou-se por não modificar esses registros, pois as datas são fornecidas diretamente pela Capes e estão formatadas corretamente. Como algumas coincidem com informações do currículo Lattes e outras não, além de haver casos sem registros, decidiu-se mantê-las inalteradas, uma vez que não é possível validá-las. Qualquer análise futura deve considerar possíveis imprecisões e interpretar os dados com cautela.

A análise segue para o campo NM_ENTIDADE_ENSINO, que registra os nomes das instituições de ensino, que contém 1.237.975 registros (total de linhas) e 686 nomes únicos. A revisão no LibreOffice Calc identificou variações na grafia, uso inconsistente de abreviaturas e diferentes formas de registrar campi. Apesar dessas variações, para preservar a integridade das integrações de dados, decidiu-se por não alterar os nomes. Esse campo serve como referência para cruzamentos com outros conjuntos de dados, e modificações poderiam comprometer a consistência das validações entre bases relacionadas aos programas de pós-graduação.

A análise segue para o campo NM_PAIS_ORIGEM_DISCENTE, que registra a nacionalidade dos discentes e contém 168 países distintos e 1.237.975 entradas (totalmente preenchido). A inspeção revelou inconsistências na grafia dos nomes dos países, como "SUIÇA" e "SUÍÇA", além de categorias como "NÃO INFORMADO", "NI" e "OUTRO", que afetam a uniformidade dos dados. Para corrigir essas discrepâncias, foi criado um *script*, ilustrado no Quadro 85, que utiliza um dicionário de valores pré-definido para padronizar os nomes dos países e salvar as alterações no DF6_discentes_BR.

Quadro 85 – Script desenvolvido para alterar os valores da coluna NM_PAIS_ORIGEM_DISCENTE

```
import pandas as pd
# Caminho do arquivo
file_path = '/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF6/DF6.csv'
# Lei o arquivo CSV
df = pd.read_csv(file_path, sep=';', encoding='utf-8')
# Dicionário com os valores a serem alterados
alteracoes={"ARMÊNIA": "ARMÊNIA", "BOLÍVIA": "BOLÍVIA", "BULGÁRIA": "BULGÁRIA", "CANADA": "CANADÁ", "COLOMBIA": "COLÔMBIA", "COREIA DO SUL": "CORÉIA DO SUL", "COSTA DO MARFIM": "COSTA DO MARFIM", "IUGUSLÁVIA": "IUGOSLÁVIA", "NAMÍBIA": "Namíbia", "NI": "NÃO INFORMADO", "NICARAGUA": "NICARÁGUA", "OUTRO": "NÃO INFORMADO", "POLONIA": "POLÔNIA", "REP. DOS CAMARÕES": "CAMARÕES", "REPÚBLICA DO CONGO": "REPÚBLICA DEMOCRÁTICA DO CONGO", "RUSSIA": "RÚSSIA", "SUIÇA": "SUÍÇA", "TIMOR LESTE": "TIMOR-LESTE",}
# Realize as alterações no campo NM_PAIS_ORIGEM_DISCENTE
df["NM_PAIS_ORIGEM_DISCENTE"] = df["NM_PAIS_ORIGEM_DISCENTE"].replace(alteracoes)
# Salve o DataFrame atualizado de volta ao arquivo CSV
df.to_csv(file_path, sep=';', encoding='utf-8', index=False)
print("Arquivo atualizado com sucesso!")
```

Fonte: Dados de pesquisa (2024).

Após a execução do *script*, o número de países únicos foi ajustado para 152. Para resolver a questão dos 482 discentes rotulados como "Não Informado", foi realizada uma pesquisa no currículo Lattes e em buscadores online para identificar os países de origem desses discentes. Esse processo permitiu a correção de 392 registros, pois para 90 registros não foram encontrados seus respectivos currículos no Lattes (1 de 2019 e 89 referentes ao conjunto 2004-2012).

Para formalizar as correções, foi desenvolvido um novo *script*, ilustrado no Quadro 86, que faz alterações no campo NM_PAIS_ORIGEM_DISCENTE utilizando um dicionário de critérios que inclui AN_BASE, AN_NASCIMENTO_DISCENTE, CD_PROGRAMA_IES e ID_PESSOA para localizar as entradas a serem corrigidas. Quando o valor no campo for "NÃO INFORMADO", ele é substituído pelo valor correspondente no campo NM_PAIS_ORIGEM_DISCENTE_NOVO. Essa abordagem aprimora a qualidade dos dados e garante maior precisão para futuras análises sobre os discentes dos programas de pós-graduação.

Quadro 86 – Script desenvolvido para alterar os valores da coluna NM_PAIS_ORIGEM_DISCENTE

```
import pandas as pd
# Caminho do arquivo
file_path = '/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF6/DF6.csv'
# Leia o arquivo CSV
df = pd.read_csv(file_path, sep=';', encoding='utf-8')
# Dicionário com os valores para busca e o novo valor para NM_PAIS_ORIGEM_DISCENTE
# Exemplo de dicionário
alteracoes = [
{
"AN_BASE": 2005,
"AN_NASCIMENTO_DISCENTE": 1983,
"CD_PROGRAMA_IES": "12345",
```

```

        "ID_PESSOA": "67890",
        "NM_PAIS_ORIGEM_DISCENTE_NOVO": "BRASIL"
    },
]
# Loop para aplicar as alterações
for alteracao in alteracoes:
    # Condição para encontrar a linha específica
    condicao = (
        (df['AN_BASE'] == alteracao['AN_BASE']) &
        (df['AN_NASCIMENTO_DISCENTE'] == alteracao['AN_NASCIMENTO_DISCENTE']) &
        (df['CD_PROGRAMA_IES'] == alteracao['CD_PROGRAMA_IES']) &
        (df['ID_PESSOA'] == alteracao['ID_PESSOA']) &
        (df['NM_PAIS_ORIGEM_DISCENTE'] == "NÃO INFORMADO")
    )
    # Atualize o valor de NM_PAIS_ORIGEM_DISCENTE para o valor fornecido
    df.loc[condicao, 'NM_PAIS_ORIGEM_DISCENTE'] = alteracao['NM_PAIS_ORIGEM_DISCENTE_NOVO']
# Salve o DataFrame atualizado de volta ao arquivo CSV
df.to_csv(file_path, sep=';', encoding='utf-8', index=False)
print("Arquivo atualizado com sucesso!")

```

Fonte: Dados de pesquisa (2024).

A atualização do campo NM_PAIS_ORIGEM_DISCENTE completou os registros, permitindo integrações com outros conjuntos de dados e a criação de mapas sobre a distribuição geográfica dos mestres e doutores. Esse aprimoramento facilita análises sobre a origem dos discentes em programas de pós-graduação e melhora a qualidade geral do DF6_discentes_BR.

A análise continua para o campo NM_ORIENTADOR, que contém 116.425 valores únicos, mas sem um identificador exclusivo para cada orientador, o que dificulta a correspondência exata dos registros. Para tratar essa limitação, foi desenvolvido um *script*, ilustrado no Quadro 87, remove acentos dos nomes e armazena os resultados em um novo arquivo CSV, preservando a versão original e a versão sem acentos no campo NM_ORIENTADOR_S_ACENTO.

Quadro 87 – Script desenvolvido para criar arquivo com o nome dos professores únicos em um csv

```

import pandas as pd
from unidecode import unidecode
# Caminho do arquivo
file_path = '/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF6/DF6.csv'
# Leia o arquivo CSV
df = pd.read_csv(file_path, sep=';', encoding='utf-8')
# Captura os nomes únicos do campo NM_ORIENTADOR
df_orientadores_unicos = pd.DataFrame(df['NM_ORIENTADOR'].unique(), columns=['NM_ORIENTADOR'])
# Crie o campo NM_ORIENTADOR_S_ACENTO removendo acentos
df_orientadores_unicos['NM_ORIENTADOR_S_ACENTO'] = df_orientadores_unicos['NM_ORIENTADOR'].apply(unidecode)
# Salve o DataFrame resultante em um novo arquivo CSV
output_file_path = '/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF6/NM_ORIENTADOR_UNICOS_S_ACENTO.csv'
df_orientadores_unicos.to_csv(output_file_path, sep=';', encoding='utf-8', index=False)
print(f"Arquivo salvo com sucesso em: {output_file_path}")

```

Fonte: Dados de pesquisa (2024).

O processamento desse *script* permite a comparação com o DF5_docentes_BR, possibilitando a recuperação de identificadores únicos para os orientadores. Além disso,

corrige inconsistências na grafia, como a diferenciação entre "SERGIO" e "SÉRGIO", garantindo a normalização dos dados para maior precisão nas integrações e análises. O campo NM_ORIENTADOR_S_ACENTO foi criado para padronizar os registros e evitar duplicações que poderiam comprometer a integridade da análise.

O arquivo gerado foi aberto no LibreOffice Calc, revelando variações na grafia que resultam em registros duplicados. Para resolver isso, foi desenvolvido um *script*, ilustrado no Quadro 89, que processa os dados do Quadro 88. Esse *script* carrega os nomes sem acento e busca correspondências no DF5_docentes_BR, atribuindo o ID_DOCENTE quando encontrado, garantindo a correta associação dos orientadores aos seus respectivos registros.

Quadro 88 – Script desenvolvido para buscar o mesmo valor entre os campos sem acentos no DataFrame DF5_docentes_BR e NM_ORIENTADOR_UNICOS_S_ACENTO.csv

```
import pandas as pd
# Caminho dos arquivos
file_orientador_path = '/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF6/NM_ORIENTADOR_UNICOS_S_ACENTO.csv'
file_docente_path = '/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF5/DF5.csv'
# Leia os arquivos CSV
df_orientador = pd.read_csv(file_orientador_path, sep=';', encoding='utf-8')
df_docente = pd.read_csv(file_docente_path, sep=';', encoding='utf-8')
# Adicionar uma nova coluna para ID_DOCENTE no df_orientador
df_orientador['ID_DOCENTE'] = None
# Iterar sobre cada linha do df_orientador e buscar o valor correspondente no df_docente
for index, row in df_orientador.iterrows():
    nm_orientador_s_acento = row['NM_ORIENTADOR_S_ACENTO']
    # Verifique se há uma correspondência em NM_DOCENTE_S_ACENTO no arquivo DF5
    match = df_docente[df_docente['NM_DOCENTE_S_ACENTO'] == nm_orientador_s_acento]
    if not match.empty:
        # Se encontrar a correspondência, pegue o ID_DOCENTE e salve na coluna correspondente
        df_orientador.at[index, 'ID_DOCENTE'] = match.iloc[0]['ID_DOCENTE']
# Salve o DataFrame atualizado de volta ao arquivo CSV
df_orientador.to_csv(file_orientador_path, sep=';', encoding='utf-8', index=False)
print(f'Arquivo {file_orientador_path} atualizado com os valores de ID_DOCENTE.")
```

Fonte: Dados de pesquisa (2024).

O *script* permitiu recuperar o ID_DOCENTE para a maioria dos docentes no DF6_discentes_BR, mas algumas inconsistências impediram a correspondência automática. Erros como inversão de nomes e sobrenomes, espaços duplos e omissões de partículas como "dos", "de" e "da" afetaram a identificação. Para corrigir essas discrepâncias, ajustes foram feitos manualmente no LibreOffice Calc, inserindo diretamente os IDs no arquivo CSV.

Apesar dos ajustes, 919 casos permaneceram sem IDs, devido à variação na grafia dos nomes ou à ausência de identificadores correspondentes. Como solução, foi adicionado o campo ID_DOCENTE no DF6_discentes_BR e o campo NM_ORIENTADOR foi renomeado para NM_DOCENTE. O processo de criação desse campo e a padronização dos nomes dos orientadores como docentes são detalhados no *script* mostrado no Quadro 89.

Quadro 89 – Script desenvolvido para inserção do campo ID_DOCENTE e alteração do campo NM_ORIENTADOR para NM_DOCENTE

```
import pandas as pd
# Caminho do arquivo
file_path = '/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF6/DF6.csv'
# Leia o arquivo CSV
df = pd.read_csv(file_path, sep=';', encoding='utf-8')
# Renomeie a coluna NM_ORIENTADOR para NM_DOCENTE
df = df.rename(columns={'NM_ORIENTADOR': 'NM_DOCENTE'})
# Adicione uma nova coluna ID_DOCENTE com valores nulos (ou outro valor padrão, se necessário)
df['ID_DOCENTE'] = None
# Salve o DataFrame atualizado no mesmo arquivo
df.to_csv(file_path, sep=';', encoding='utf-8', index=False)
print(f"Coluna 'NM_ORIENTADOR' renomeada para 'NM_DOCENTE' e a coluna 'ID_DOCENTE' foi adicionada.")
print(f"Arquivo '{file_path}' atualizado com sucesso!")
```

Fonte: Dados de pesquisa (2024).

Após adicionar a coluna ID_DOCENTE ao DF6_discentes_BR, foi necessário integrar os IDs recuperados do DF5_docentes_BR. Para isso, foi desenvolvido o *script* do Quadro 90, que busca correspondências entre os nomes dos orientadores no DF6_discentes_BR e os registros no DF5_docentes_BR. O *script* lê o DF5_docentes_BR e compara os nomes com aqueles do arquivo NM_ORIENTADOR_UNICOS_S_ACENTO_EDITADO.csv, que contém os nomes normalizados dos orientadores. Quando uma correspondência é encontrada, o ID_DOCENTE associado no DF5_docentes_BR é inserido no campo correspondente do DF6_discentes_BR. Esse procedimento assegura a consistência na identificação dos docentes entre o DF6_discentes_BR e o DF5_docentes_BR, permitindo análises integradas e a rastreabilidade das informações.

Quadro 90 – Script desenvolvido para inserir os valores dos ID_DOCENTE no DF6_discentes_BR

```
import pandas as pd
from unidecode import unidecode
# Caminhos dos arquivos
df6_path = '/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF6/DF6.csv'
orientador_path =
'/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF6/NM_ORIENTADOR_UNICOS_S_ACENTO_EDITADO.csv'
# Leia os arquivos CSV
df6 = pd.read_csv(df6_path, sep=';', encoding='utf-8')
orientador_df = pd.read_csv(orientador_path, sep=';', encoding='utf-8')
# Função para remover acentos e caracteres especiais
def remover_acentos(texto):
    return unidecode(texto) if isinstance(texto, str) else texto
# Iterar sobre cada linha do df6 e buscar o valor correspondente no NM_ORIENTADOR_UNICOS_S_ACENTO_EDITADO
for index, row in df6.iterrows():
    nm_docente = row['NM_DOCENTE']
    # Remover acentos e caracteres especiais de NM_DOCENTE
    nm_docente_s_acento = remover_acentos(nm_docente)
    # Verifique se há uma correspondência em NM_ORIENTADOR_S_ACENTO no arquivo
    NM_ORIENTADOR_UNICOS_S_ACENTO_EDITADO
    match = orientador_df[orientador_df['NM_ORIENTADOR_S_ACENTO'] == nm_docente_s_acento]
    if not match.empty:
        # Se encontrar a correspondência, pegue o ID_DOCENTE e salve no campo correspondente de df6
        df6.at[index, 'ID_DOCENTE'] = match.iloc[0]['ID_DOCENTE']
# Salve o DataFrame atualizado de volta ao arquivo CSV
```

```
df6.to_csv(df6_path, sep=';', encoding='utf-8', index=False)
print(f'Arquivo '{df6_path}' atualizado com os valores de ID_DOCENTE.")
```

Fonte: Dados de pesquisa (2024).

Após a execução do *script*, o campo ID_DOCENTE no DF6_discentes_BR foi preenchido com os dados extraídos do DF5_docentes_BR. No entanto, 2.541 registros permanecem com o valor "NÃO INFORMADO" e 1.122 linhas não tiveram correspondência nos dados fornecidos pela Capes. A revisão manual identificou variações na grafia e omissões de partes dos nomes, dificultando a correspondência automática. Consultas ao currículo Lattes e buscas em plataformas online foram realizadas, mas algumas inconsistências persistem, indicando falhas nos registros originais. Diante disso, optou-se por manter o campo ID_DOCENTE como está, com necessidade de uma revisão futura mais detalhada. Essa decisão destaca a complexidade de trabalhar com grandes conjuntos de dados acadêmicos e a importância de um processo cuidadoso na curadoria de dados para pesquisa.

A análise prossegue com o campo DS_FAIXA_ETARIA, que categoriza os discentes em 12 faixas etárias, conforme apresentado na Tabela 31.

Tabela 31 – Valores encontrados no campo DS_FAIXA_ETARIA

Valor	Quantidade
19 OU MENOS	38
20 A 24 ANOS	37054
25 A 29 ANOS	420321
30 A 34 ANOS	320012
35 A 39 ANOS	181967
40 A 44 ANOS	115439
45 A 49 ANOS	76120
50 A 54 ANOS	47795
55 A 59 ANOS	25366
60 A 64 ANOS	9927
65 A 69 ANOS	3041
70 OU MAIS	895

Fonte: Dados de pesquisa (2024).

Entre os registros, 38 discentes são classificados como "19 ou menos", incluindo cinco doutorandos com tempo médio de conclusão de 44 meses, o que levanta dúvidas sobre a precisão desses dados. A ausência de acesso aos dados de nascimento impede a verificação

direta dessas idades, tornando necessária uma interpretação cautelosa das faixas etárias. Essas discrepâncias podem impactar análises demográficas e avaliações sobre o perfil dos discentes nos programas de pós-graduação no Brasil.

A análise no DF6_discentes_BR segue para o campo NM_NIVEL_CONCLUSAO_DISCENTE, que indica o grau acadêmico concluído pelo discente. Conforme apresentado na Tabela 32, o campo inclui a categoria "NI" (não informado), refletindo lacunas nos dados fornecidos pela Capes. A presença desse valor pode comprometer análises sobre o perfil e a taxa de conclusão dos programas, destacando a necessidade de avaliar a qualidade dos registros e possíveis estratégias para minimizar a incompletude dos dados.

Tabela 32 – Valores encontrados no campo NM_NIVEL_CONCLUSAO_DISCENTE

Valor	Quantidade
MESTRADO	788907
DOUTORADO	302638
MESTRADO PROFISSIONAL	143889
NI	2424
DOUTORADO PROFISSIONAL	117

Fonte: Dados de pesquisa (2024).

A presença dos valores "NI" pode comprometer análises sobre o perfil e a taxa de conclusão dos programas, destacando a necessidade de avaliar a qualidade dos registros e possíveis estratégias para minimizar a incompletude dos dados.

Continuando a análise do DF6_discentes_BR, identificaram-se duplicações de nomes de discentes, resultando em campos vazios no NM_NIVEL_CONCLUSAO_DISCENTE e no NM_DOCENTE. A duplicidade está relacionada à variável NM_SITUACAO_DISCENTE, que classifica os discentes como "TITULADO" ou "MUDANÇA DE NÍVEL SEM DEFESA" para indicar a conclusão do curso. No entanto, em alguns casos, a mudança de nível registrada não corresponde ao ano real de conclusão. Para corrigir essa inconsistência, os registros foram revisados e as duplicações removidas, garantindo que cada entrada representasse corretamente a trajetória acadêmica do discente.

A análise no DF6_discentes_BR identificou 155 países no campo NM_PAIS_ORIGEM_DISCENTE, com correção da grafia de "GUINÉ BISSAU" para "GUINÉ-BISSAU", conforme a nomenclatura geográfica padrão. Além disso, 96 registros estavam como "NÃO INFORMADO". Após investigações no currículo Lattes e buscas

online, foi possível corrigir seis registros, reduzindo esse número para 90. Essa ação contribui para a precisão do dataset e melhora a qualidade das análises futuras sobre a origem geográfica dos discentes.

O campo QT_MES_TITULACAO, que registra o tempo (em meses) entre matrícula e conclusão do curso, apresentou discrepâncias. Foram identificados-se 17 registros com valores negativos, indicando possíveis erros nas datas, 3.121 registros com valor zero, sugerindo falha no cadastramento, e 9.756 registros inferiores a 12 meses, abaixo do tempo mínimo esperado para um Mestrado. Considerando esses problemas, esses valores requerem uma análise mais detalhada para verificar possíveis falhas no cadastramento das datas. Sem fontes externas para validação, essas inconsistências podem impactar a integridade dos dados para análises futuras.

Os ajustes no DF6_discentes_BR incluíram a padronização de datas, correção de grafias, ajustes nos campos de nacionalidade, nível de conclusão e faixa etária. Essas modificações aprimoram a precisão dos dados, permitindo análises estatísticas mais consistentes e facilitando a integração com outros conjuntos de dados, tornando o DF6_discentes_BR uma base confiável para pesquisas e formulação de políticas educacionais baseadas em dados.

Após a correção dos dados de programas, docentes e discentes, a atenção voltou-se para o aprimoramento do DF3_trabalhos_BR, o DataFrame central desta pesquisa.

4.5.1.5.5 Fase 5 - Análise exploratória profunda dos dados no DF3_trabalhos_BR

Com os dados adequadamente preparados, é conduzida uma análise exploratória, procurando tendências, características e peculiaridades nos dados, podendo envolver testes estatísticos, modelagem e outras técnicas avançadas para extrair informações relevantes dos dados e chegar à visualização de forma gráfica dos dados (Grandvalet; Bengio, 2004; Witten *et al.*, 2005; Huang *et al.*, 2015).

Como mostrado na Fase 3, na Tabela 08, o DF3_trabalhos_BR apresenta lacunas em campos como CD_GRANDE_AREA_CONHECIMENTO, NM_GRANDE_AREA_CONHECIMENTO, NM_REGIAO e SG_UF_IES, que precisam ser avaliados quanto à possibilidade de preenchimento. Além disso, a integridade dos dados nas 24 colunas do DataFrame foi verificada, garantindo precisão e consistência.

O DF3_trabalhos_BR é composto por 24 colunas, sendo ANO_BASE a primeira, indicando o ano de conclusão dos discentes. Esse campo contém 1.232.870 registros

completos, abrangendo 19 anos (2004 a 2022), conforme os dados da CAPES, permitindo análises temporais da pós-graduação no Brasil.

O próximo foco do ajuste no DF3_trabalhos_BR concentraram-se inicialmente nos dois campos: o código e o nome da grande área do conhecimento. O campo do código contém 1.232.854 entradas, enquanto o nome possui 1.232.849, ambos com 9 valores únicos, correspondendo às nove grandes áreas definidas pela CAPES. Para garantir consistência, esses campos foram tratados simultaneamente, associando corretamente cada código ao respectivo nome.

Para auxiliar a verificação e correção dos dados, foi desenvolvido o *script* do Quadro 91, que processa o arquivo DF3_com_indice.csv. O *script* solicita a coluna a ser analisada, identifica registros vazios (seja NaN ou *strings* em branco) e exporta os resultados para um novo arquivo CSV, facilitando a revisão e ajustes nos dados.

Quadro 91 – Script para exportar as linhas vazias de uma determinada coluna

```
import pandas as pd
# Caminho do arquivo
file_path = '/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF3/DF3_com_indice.csv'
# Leia o arquivo CSV
df = pd.read_csv(file_path, sep=';', encoding='utf-8', low_memory=False)
# Solicite ao usuário o nome da coluna para verificar valores vazios
column_name = input("Digite o nome da coluna que deseja verificar por valores vazios: ")
# Verifique se a coluna existe no DataFrame
if column_name not in df.columns:
    print(f"A coluna '{column_name}' não existe no arquivo.")
else:
    # Inspeção inicial dos valores únicos na coluna
    unique_values = df[column_name].dropna().unique()
    print(f"Valores únicos na coluna '{column_name}' (parcial): {unique_values[:20]}")
    # Identifique valores vazios de acordo com o tipo de dado
    if pd.api.types.is_numeric_dtype(df[column_name]):
        # Para colunas numéricas, considere apenas valores NaN
        empty_rows = df[df[column_name].isna()]
    elif pd.api.types.is_string_dtype(df[column_name]):
        # Para colunas de texto, verifique NaN, strings vazias, e espaços
        empty_rows = df[
            df[column_name].isna() |
            (df[column_name] == "") |
            (df[column_name].str.strip() == "")
        ]
    else:
        # Para outros tipos de dados, considere apenas valores NaN
        empty_rows = df[df[column_name].isna()]
    # Exiba o número total de linhas vazias encontradas
    print(f"Total de linhas vazias na coluna '{column_name}': {empty_rows.shape[0]}")
    # Exibir as primeiras linhas como exemplo
    if not empty_rows.empty:
        print(f"Exemplo de linhas vazias:\n{empty_rows.head()}")
    # Caminho para salvar o arquivo de saída
    output_file_path = f'/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF3/{column_name}_linhas_vazias.csv'
    # Salve as linhas vazias em um novo arquivo CSV
    empty_rows.to_csv(output_file_path, sep=';', encoding='utf-8', index=False)
    print(f"As linhas vazias foram salvas no arquivo: {output_file_path}")
```

Fonte: Dados de pesquisa (2024).

Após a execução inicial do *script*, foram gerados dois arquivos CSV, um contendo os códigos e outro com os nomes das grandes áreas de conhecimento. A análise desses arquivos revelou que os registros sem código também estavam sem nomes, permitindo concentrar as correções no arquivo de nomes. Além disso, foi possível atualizar outros campos vazios, como resumos e nomes dos orientadores. Para aplicar as correções no DF3_trabalhos_BR, foi criado o *script* ilustrado no Quadro 92, que atualiza os dados com base nos índices das colunas indicadas no arquivo CSV.

Quadro 92 – Script para atualizar linhas no arquivo DF3_trabalhos_BR com base em um csv

```
import pandas as pd
# Caminho dos arquivos
file_to_update_path = '/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF3/DF3_com_indice.csv'
file_with_updates_path =
'/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF3/NM_GRANDE_AREA_CONHECIMENTO_linhas_ok.csv'
# Leia os arquivos CSV
df_to_update = pd.read_csv(file_to_update_path, sep=';', encoding='utf-8', low_memory=False)
df_with_updates = pd.read_csv(file_with_updates_path, sep=';', encoding='utf-8')
# Certifique-se de que a coluna 'indice' está presente em ambos os DataFrames
if 'indice' not in df_to_update.columns or 'indice' not in df_with_updates.columns:
    print("A coluna 'indice' não está presente em um dos arquivos.")
else:
    # Realize a atualização dos dados no DataFrame principal
    df_to_update = df_to_update.set_index('indice')
    df_with_updates = df_with_updates.set_index('indice')
    # Atualizar os dados correspondentes
    df_to_update.update(df_with_updates)
    # Voltar o índice para coluna padrão
    df_to_update.reset_index(inplace=True)
    # Salvar o arquivo atualizado
    df_to_update.to_csv(file_to_update_path, sep=';', encoding='utf-8', index=False)
    print(f"Arquivo atualizado com sucesso: {file_to_update_path}")
```

Fonte: Dados de pesquisa (2024).

No DF3_trabalhos_BR, os campos CD_AREA_CONHECIMENTO e NM_AREA_CONHECIMENTO registram, respectivamente, o código e o nome da área de conhecimento associada a cada programa de pós-graduação. O campo de código está completo, enquanto o de nome apresentou três células vazias. Para corrigir essas lacunas, foi utilizado o *script* ilustrado no Quadro 88, que identificou e ajustou os registros. Após a correção, o script no Quadro 89 foi utilizado para atualizar os dados no DF3_trabalhos_BR, garantindo a correspondência entre ambos os campos.

O campo NM_AREA_AVALIACAO apresentou sete linhas vazias, o que exigiu correção. Com o script do Quadro 88, foi gerado um arquivo CSV para visualizar essas ausências e outras lacunas, como em resumos e identificadores de orientadores. Para corrigir as lacunas, foi seguido o procedimento do Quadro 89 para atualizar o DF3_trabalhos_BR. Além das lacunas, foram identificadas variações na nomenclatura, como “FILOSOFIA / TEOLOGIA: subcomissão FILOSOFIA” e “FILOSOFIA/TEOLOGIA: SUBCOMISSÃO

FILOSOFIA”, que foram padronizadas conforme o portal da Capes, resultando em 52 denominações únicas para o campo NM_AREA_AVALIACAO.

Os campos CD_PROGRAMA_IES e NM_PROGRAMA_IES, que registram o código e o nome dos programas, estavam completos, sem valores nulos. A análise mostrou 5.047 códigos distintos e 2.631 nomes de programas. Como múltiplos códigos podem estar associados a um único nome — como ocorre nos programas de Ciência da Informação, que possuem 16 códigos distintos — decidiu-se manter os campos inalterados para preservar a integridade dos dados.

Na revisão dos campos SG_ENTIDADE_ENSINO e NM_ENTIDADE_ENSINO, que representam respectivamente a sigla e o nome das IES, identificaram-se inconsistências. Algumas siglas estavam associadas a múltiplas instituições, contrariando sua função de identificador único. Para corrigir essas inconsistências, foi utilizado um *script*, descrito no Quadro 93, para tratar casos como o da FUNDACAO VISCONDE DE CAIRU, inicialmente registrada com a sigla FVC.

Quadro 93 – Script desenvolvido para ajustar valores de uma coluna filtrado por duas colunas

```
import pandas as pd
# Condições para identificar as linhas a serem alteradas
condition = (df['NM_ENTIDADE_ENSINO'] == 'FUNDACAO VISCONDE DE CAIRU') & (df['SG_ENTIDADE_ENSINO'] == 'FVC'))
# Contar as linhas antes da alteração
rows_to_update = condition.sum()
# Atualizar as linhas que atendem às condições
df.loc[condition, 'SG_ENTIDADE_ENSINO'] = 'FVCAIRU'
# Contar as linhas após a alteração
rows_updated = (df['SG_ENTIDADE_ENSINO'] == 'FVCAIRU').sum()
# Exibir resumo das alterações
print(f"Linhas identificadas para alteração: {rows_to_update}")
print(f"Linhas atualizadas para 'CAIRU': {rows_updated}")
# Salvar o DataFrame atualizado
file_path = '/content/drive/MyDrive/CAPES/datasets/DF3/DF3.csv'
df.to_csv(file_path, sep=',', encoding='utf-8', index=False)
print(f"Arquivo atualizado salvo em: {file_path}")
```

Fonte: Dados de pesquisa (2024).

O *script* aplicou uma substituição condicional, corrigindo a sigla para FVCAIRU nas linhas onde a condição NM_ENTIDADE_ENSINO == 'FUNDACAO VISCONDE DE CAIRU' e SG_ENTIDADE_ENSINO == 'FVC' era atendida, garantindo a consistência na identificação das instituições.

Além disso, foi necessária a padronização dos nomes das instituições, que apresentavam variações nas abreviações e grafias. Para esse ajuste, foi desenvolvido um *script*, ilustrado no Quadro 94, que realizou a substituição sistemática dos nomes. O *script* recebeu como entrada o nome da coluna a ser ajustada, os valores incorretos identificados e os

respectivos nomes padronizados, para garantir maior uniformidade nos registros institucionais e, assim, facilitar as análises e integrações futuras.

Quadro 94 – Script desenvolvido para ajustar valores de uma coluna

```
import pandas as pd
# Nome da coluna onde será realizada a substituição
column_name = "SG_ENTIDADE_ENSINO"
# Valores a serem alterados
values_to_replace = [
    "UFPE", "UPE", "FESP/UPE",
]
# Novo valor a ser inserido
new_value = "UFPE"
# Caminho do arquivo para salvar os dados atualizados
file_path = "/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF3/DF3.csv"
# Verifique se a coluna existe no DataFrame
if column_name not in df.columns:
    print(f'A coluna '{column_name}' não existe no DataFrame.")
else:
    # Contar as ocorrências dos valores antes da modificação
    occurrences_before = df[column_name].isin(values_to_replace).sum()
    # Realizar a substituição diretamente no DataFrame
    df[column_name] = df[column_name].replace(values_to_replace, new_value)
    # Contar as ocorrências do novo valor após a modificação
    occurrences_after = (df[column_name] == new_value).sum()
    # Exibir resumo das alterações
    print(f'Substituição concluída na coluna '{column_name}'.')
    print(f'Total de valores alterados: {occurrences_before}")
    print(f'Total de ocorrências do novo valor '{new_value}': {occurrences_after}")
    # Salvar o DataFrame atualizado no arquivo especificado
    df.to_csv(file_path, sep=';', encoding='utf-8', index=False)
    print(f'Arquivo atualizado salvo em: {file_path}")
```

Fonte: Dados de pesquisa (2024).

Após as atualizações, o número de siglas e nomes de instituições foi consolidado em 491 identificadores únicos, garantindo maior integridade ao DF3_trabalhos_BR e facilitando análises futuras. Essas correções são necessárias para garantir a confiabilidade e a integridade do DF3_trabalhos_BR, facilitando análises futuras mais precisas.

Em seguida, a revisão do campo NM_REGIAO, que registra a região geográfica das IES, identificou cinco valores únicos, correspondentes às cinco regiões do Brasil. Para validar esses dados, foi utilizado o *script* do Quadro 95.

Quadro 95 – Script desenvolvido para filtrar valores por duas colunas

```
import pandas as pd
# Solicite ao usuário o nome das colunas
unique_column = input("Digite o nome da coluna para filtrar os valores únicos: ")
associated_column = input("Digite o nome da coluna onde os dados associados serão apresentados: ")
# Verifique se as colunas existem no DataFrame
if unique_column not in df.columns or associated_column not in df.columns:
    print(f"Uma das colunas especificadas ('{unique_column}' ou '{associated_column}') não existe no DataFrame.")
else:
    # Filtrar os valores únicos e os dados correspondentes
    grouped_data = df.groupby(unique_column)[associated_column].apply(lambda x: ', '.join(map(str, x.unique()))).reset_index()
    grouped_data.columns = [unique_column, f"Valores em '{associated_column}'"]
    # Exibir os resultados
    print(f"Dados agrupados por '{unique_column}' com valores de '{associated_column}':")
    print(grouped_data)
    # Salvar o resultado em um arquivo CSV
    output_file_path = f'content/drive/MyDrive/CAPES/datasets/dados_abertos/DF3/{unique_column} com {associated_column}.csv'
    grouped_data.to_csv(output_file_path, sep=',', encoding='utf-8', index=False)
    print(f"Resultados salvos em: {output_file_path}")
```

Fonte: Dados de pesquisa (2024).

O processamento do *script* permitiu uma revisão detalhada dos dados, que comparou as informações com o campo SG_UF_IES, assegurando a coerência entre estados e regiões, conforme mostrado na Tabela 33.

Tabela 33 – Regiões e UF presentes no DF3_trabalhos_BR

Nome da região	Nomes das UF
CENTRO-OESTE	GO, DF, MS, MT
NORDESTE	CE, PB, PE, BA, RN, SE, MA, PI, AL
NORTE	AM, PA, AC, RO, TO, RR, AP
SUDESTE	RJ, SP, MG, ES
SUL	PR, SC, RS

Fonte: Dados de pesquisa (2024).

A verificação confirmou a consistência dos campos NM_REGIAO e SG_UF_IES, sem valores ausentes ou discrepantes, possibilitando seu uso em análises geográficas, como mapas de calor e densidade sobre a distribuição dos programas de pós-graduação no Brasil.

A análise do DF3_trabalhos_BR seguiu para os campos ID_ADD_PRODUCAO_INTELECTUAL e ID_PRODUCAO_INTELECTUAL, que identificam teses e dissertações na base da CAPES. O primeiro corresponde a um identificador único vinculado a um ano base e programa, enquanto o segundo é um identificador permanente que permite consulta direta dos trabalhos concluídos a partir de 2013 na plataforma Sucupira. Ambos os campos estão completos, sem valores nulos, garantindo a rastreabilidade e integridade dos registros acadêmicos.

Passando para a análise do DF3_trabalhos_BR, nos campos ID_GRAU_ACADEMICO e NM_GRAU_ACADEMICO, que identificam o grau acadêmico do discente ao final do ano base, verificou-se que ambos estão completamente preenchidos, o que facilita futuras revisões. O *script* do Quadro 92 validou a correspondência entre códigos e nomes, revelando uma inconsistência no grau de Mestrado, registrado como "Mestrado" e "MESTRADO". Para padronizar os dados, foi aplicado o *script* do Quadro 90, reduzindo a variedade de nomes de sete para quatro, conforme os padrões do Catálogo de Teses e Dissertações da CAPES, garantindo a consistência das nomenclaturas, como mostrado na Tabela 34.

Tabela 34 – Identificação e o nome do Grau acadêmico presentes no DF3_trabalhos_BR

Identificação	Nome
27	MESTRADO
28	DOUTORADO
30	MESTRADO PROFISSIONAL
31	DOUTORADO PROFISSIONAL

Fonte: Dados de pesquisa (2024).

No refinamento do DF3_trabalhos_BR, o campo NM_IDIOMA, que registra o idioma do trabalho acadêmico, foi analisado. Identificaram-se sete registros nulos, corrigidos conforme os procedimentos descritos nos Quadros 88 e 89. Além disso, variações na grafia de "Português" apareceram como "PORTUGUES", "Português", "PORTUGUÊS" e "português". Para padronizar esses registros, foi aplicado o *script* do Quadro 91, consolidando os idiomas em 47 categorias distintas e garantindo a uniformidade dos dados.

Na atualização do DF3_trabalhos_BR, foram analisados os campos ID_PESSOA_ORIENTADOR, que identifica o orientador, e NM_ORIENTADOR, que registra o nome do orientador. O campo de identificação apresentou 159 valores nulos, enquanto o campo de nome teve 20. Houve uma discrepância de 139 registros entre os valores únicos dos dois campos, sugerindo inconsistências na correspondência entre identificação e nome. Para resolver essas questões, foram aplicados os *scripts* mostrados nos Quadros 88 e 89 para tratar os valores nulos e o Quadro 92 para exportar e analisar os dados filtrados, garantindo a consistência e alinhamento dos registros.

Os trabalhos acadêmicos identificados abrangem os anos de 2004 a 2020. Com consultas aos currículos Lattes, foi possível recuperar nomes dos orientadores a partir dos registros dos discentes. O *script* do Quadro 89 foi utilizado para inserir essas informações no

DF3_trabalhos_BR. No entanto, observou-se que, antes de 2013, o código de identificação dos orientadores era derivado do CPF, exibido parcialmente no formato "(911*****800)". Para ajustar os dados, é adicionada uma nova coluna ao DF3_trabalhos_BR com os nomes dos orientadores sem acentuação, usando o *script* do Quadro 96, seguindo a abordagem aplicada nos DF5_docentes_BR e DF6_discentes_BR. Esse passo é necessário para uniformizar os dados. Em seguida, é realizada uma consulta ao DF5_docentes_BR, que contém informações detalhadas sobre os orientadores, para avaliar a possibilidade de unificar os IDs nos dois DataFrames. Essa integração facilitará o cruzamento de dados e aumentará a consistência no conjunto de informações.

Quadro 96 – Script desenvolvido adicionar coluna NM_ORIENTADOR_S_ACENTO

```
!pip install unicode
import pandas as pd
import unicode
# Caminho para o arquivo CSV
file_path = '/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF3/DF3.csv'
# Carregar o DataFrame
df = pd.read_csv(file_path, sep=';', encoding='utf-8')
# Função para remover acentos e caracteres especiais
def remover_acentos(texto):
    if pd.isna(texto): # Verificar se o valor é NaN
        return texto
    return unicode.unidecode(str(texto))
# Criar a nova coluna 'NM_ORIENTADOR_S_ACENTO' aplicando a função para remover acentos
df['NM_ORIENTADOR_S_ACENTO'] = df['NM_ORIENTADOR'].apply(remover_acentos)
# Reordenar as colunas para colocar 'NM_ORIENTADOR_S_ACENTO' logo após 'NM_ORIENTADOR'
colunas = df.columns.tolist()
indice_nm_orientador = colunas.index('NM_ORIENTADOR')
colunas_nova_ordem = (
    colunas[:indice_nm_orientador + 1] +
    ['NM_ORIENTADOR_S_ACENTO'] +
    colunas[indice_nm_orientador + 1:]
)
df = df[colunas_nova_ordem]
# Salvar o DataFrame atualizado no mesmo arquivo
df.to_csv(file_path, sep=';', encoding='utf-8', index=False)
print("Coluna 'NM_ORIENTADOR_S_ACENTO' adicionada e valores atualizados com sucesso!")
```

Fonte: Dados de pesquisa (2024).

Com a inclusão da coluna contendo os nomes dos orientadores sem acentuação, foram identificadas inconsistências nos registros. A abertura do arquivo no LibreOffice Calc permitiu visualizar variações na grafia de nomes que não eram detectáveis apenas por programação. Foram observadas discrepâncias em registros com o mesmo código, diferenças na acentuação, presença ou ausência de apóstrofes e trocas de letras. Após os ajustes nos nomes, foi criado um *script*, mostrado no Quadro 97, para integrar os códigos dos docentes do DF5_docentes_BR ao DF3_trabalhos_BR, aprimorando a identificação dos orientadores e reduzindo divergências, facilitando futuras análises e integrações de dados.

Quadro 97 – Script desenvolvido adicionar busca de código do docente entre dataframes

```
import pandas as pd
from tqdm import tqdm
import os

# Caminhos dos arquivos
df3_path = "/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF3/DF3.csv"
df5_path = "/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF3/DF5_fatiado_filtrado.csv"
# Obter tamanho do DF3 para estimar progresso
total_size = os.path.getsize(df3_path)
# Carregar DF5 apenas com as colunas necessárias para o merge
print("Carregando DF5...")
df5 = pd.read_csv(df5_path, sep=";", encoding="utf-8", usecols=["NM_DOCENTE_S_ACENTO", "CD_PROGRAMA",
"ID_DOCENTE"])
# Renomear colunas do DF5 para facilitar a fusão
df5.rename(columns={"NM_DOCENTE_S_ACENTO": "NM_ORIENTADOR_S_ACENTO"}, inplace=True)
# Criar um dicionário a partir de DF5 para evitar `merge` direto
df5_dict = df5.set_index(["NM_ORIENTADOR_S_ACENTO", "CD_PROGRAMA"])[["ID_DOCENTE"]].to_dict()
# Ler DF3 em chunks para evitar sobrecarga de memória
chunk_size = 100000
df3_chunks = []
print(f"Carregando DF3 em chunks de {chunk_size} linhas...")
with pd.read_csv(df3_path, sep=";", encoding="utf-8", chunksize=chunk_size) as reader:
    for chunk in tqdm(reader, total=total_size // (chunk_size * 200)):
        # Atualiza ID_PESSOA_ORIENTADOR sem remover valores existentes!
        chunk["ID_PESSOA_ORIENTADOR"] = chunk.apply(
            lambda row: df5_dict.get((row["NM_ORIENTADOR_S_ACENTO"], row["CD_PROGRAMA"])),
            row["ID_PESSOA_ORIENTADOR"]),
            axis=1
        )
        df3_chunks.append(chunk)
# Concatenar os chunks atualizados
df3 = pd.concat(df3_chunks, ignore_index=True)
# Salvar o DF3 atualizado mantendo todas as colunas originais
print("Salvando arquivo atualizado...")
df3.to_csv(df3_path, sep=";", encoding="utf-8", index=False)
print("Processo concluído com sucesso!")
```

Fonte: Dados de pesquisa (2024).

Após a execução do *script*, o DF3_trabalhos_BR foi revisado em planilha para confirmar a aplicação das atualizações. O campo NM_ORIENTADOR_S_ACENTO registrou 101.880 valores distintos, enquanto ID_DOCENTE apresentou 102.069, refletindo a presença de homônimos. O ID_DOCENTE foi essencial para diferenciar os registros, garantindo maior precisão na identificação dos orientadores.

Os campos CD_PROGRAMA e NM_PROGRAMA que representam, respectivamente, o código e o nome dos programas de pós-graduação, foram analisados e não apresentaram valores nulos. Foram identificados 5.047 códigos únicos e 2.631 nomes únicos. A abertura do DF3_trabalhos_BR em formato de planilha permitiu inspecionar visualmente os códigos e nomes, revelando que CD_PROGRAMA estava armazenado como *string* em vez de número inteiro. Além disso, alguns códigos estavam associados a múltiplos nomes de programas, o que foi corrigido com o *script* do Quadro 98. Também foi identificada a existência de programas em rede, onde um mesmo código e nome estavam vinculados a diferentes instituições. Essas questões foram ajustadas para garantir maior consistência nas informações registradas no DF3_trabalhos_BR.

Quadro 98 – Script atualizar os nomes dos programas de pós-graduação pelos códigos

```
import pandas as pd
# Caminho do arquivo
df3_path = "/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF3/DF3.csv"
# Dicionário com os valores a serem alterados (CD_PROGRAMA: NM_PROGRAMA_NOVO)
alteracoes = {
    "24009016008P0": "Novo Programa A",
    "15001016090P1": "Novo Programa B",
    "15001016091P8": "Novo Programa C",
    # Adicione mais valores conforme necessário
}
# Carregar o arquivo CSV
print("Carregando DF3...")
df3 = pd.read_csv(df3_path, sep=";", encoding="utf-8", low_memory=False)
# Atualizar os valores no DataFrame
df3["NM_PROGRAMA"] = df3["CD_PROGRAMA"].map(lambda x: alteracoes[x] if x in alteracoes else
df3.loc[df3["CD_PROGRAMA"] == x, "NM_PROGRAMA"].values[0])
# Salvar o arquivo atualizado
df3.to_csv(df3_path, sep=";", encoding="utf-8", index=False)
print(f"Arquivo atualizado com sucesso: {df3_path}")
```

Fonte: Dados de pesquisa (2024).

Uma consulta à Plataforma Sucupira permitiu verificar e corrigir os nomes dos programas no DF3_trabalhos_BR. Alguns estavam registrados como linhas de pesquisa, apresentavam variações de maiúsculas e minúsculas ou continham erros ortográficos. Após os ajustes, o campo NM_PROGRAMA passou a conter 4.853 nomes distintos, enquanto CD_PROGRAMA manteve 5.047 códigos únicos, garantindo alinhamento com os dados oficiais.

Em seguida, os campos ID_DISCENTE e NM_DISCENTE foram analisados, sendo verificados 1.232.870 registros para o primeiro campo e 926.864 registros para o segundo, evidenciando discrepâncias. Foram identificados 736.293 valores únicos para ID_DISCENTE e 1.094.135 para NM_DISCENTE, indicando duplicações e variações na grafia. A padronização do campo NM_DISCENTE apresenta desafios, pois os nomes podem sofrer alterações ao longo do tempo, como mudanças por matrimônio, além de conter abreviações, erros de escrita e variações ortográficas.

Uma análise em planilha permitiu visualizar melhor essas variações. Consultas à Plataforma Lattes e a mecanismos de busca mostraram que alguns discentes não possuem registros disponíveis, dificultando a validação externa. Diante dessas limitações, optou-se por manter NM_DISCENTE apenas para referência e validação futura, sem alterações no DF3_trabalhos_BR.

Seguindo, é analisado o campo das palavras-chave, que contém 1.232.692 registros, porém não está completamente preenchido, apresentando 1.216.796 valores únicos. Inicialmente, essa diferença poderia indicar que cerca de 16 mil trabalhos compartilham as

mesmas palavras-chave, mas uma análise mais detalhada revelou que a inconsistência se deve à ausência de um padrão na forma como as palavras-chave foram registradas.

Ao abrir os dados desse campo em planilha, observou-se o uso de diferentes delimitadores, como ;, ., ,, - e números, além de variação na quantidade de palavras-chave por trabalho, desde um único termo até frases inteiras, incluindo trechos do próprio resumo. Outro fator observado foi a presença de mais de 20 mil registros contendo menos de 15 caracteres, o que compromete a padronização do campo. Diante dessas inconsistências, decidiu-se por não utilizar esse campo na análise principal, mantendo-o no dataframe apenas para eventuais consultas futuras.

Os campos NM_PRODUCAO e DS_RESUMO, que registram o título e o resumo dos trabalhos, foram analisados no DF3_trabalhos_BR. O campo NM_PRODUCAO possui 1.232.870 registros, enquanto DS_RESUMO contém 1.232.778, indicando uma pequena discrepância. Foram identificados 1.230.782 valores únicos para títulos e 1.221.370 para resumos, com 9.412 registros distintos entre ambos. Ao abrir os dados em planilha, observou-se que alguns trabalhos possuem resumo sem título e vice-versa. Além disso, mais de 16 mil registros no campo DS_RESUMO possuem menos de 150 caracteres, apresentando apenas palavras isoladas ou caracteres soltos, o que compromete a qualidade da informação. Diante dessas limitações, não são feitos ajustes nesses campos. Quando necessário, um recorte por grande área ou área do conhecimento permitirá uma verificação mais detalhada.

No processo de ajuste do DF3_trabalhos_BR, foram realizadas correções para garantir consistência e integridade dos dados. Lacunas foram preenchidas, nomes e siglas de instituições foram padronizados, e scripts foram desenvolvidos para corrigir registros ausentes e inconsistências em identificadores de discentes, docentes e programas. Campos como palavras-chave e resumos foram analisados, mas não sofreram ajustes, sendo mantidos para referência futura.

Concluídos esses procedimentos, inicia-se a Etapa 6, que trata do recorte dos dados.

4.5.1.6 Etapa 6 – Recorte de dados

Nesta etapa, a análise concentra-se exclusivamente nos dados da Ciência da Informação. Inicialmente, a intenção era consolidar todas as informações em um único dataframe, facilitando a extração e análise dos dados em um ambiente unificado. No entanto, essa fusão não foi viável, pois há variáveis presentes em alguns dataframes que não estão disponíveis em outros, impossibilitando a integração sem perda de informações.

Diante dessa limitação, a estratégia adotada foi manter os quatro dataframes separados - DF3_trabalhos_CI, DF4_programas_CI, DF5_docentes_CI e DF6_discentes_CI -, garantindo que cada um seja analisado individualmente. O recorte é feito com base no campo da área de conhecimento, garantindo a seleção apenas dos registros relevantes para as análises subsequentes. A partir dessas análises, é possível compreender a distribuição, evolução e características da Ciência da Informação dentro do contexto da pós-graduação *stricto sensu*.

As subseções a seguir detalham os procedimentos para o recorte dos dados, a saber: Fase 1 - Recorte dos dados da Ciência da Informação no DF3_trabalhos_CI, Fase 2 - Recorte dos dados da Ciência da Informação no DF4_programas_CI, Fase 3 - Recorte dos dados da Ciência da Informação no DF5_docentes_CI e Fase 4 - Recorte dos dados da Ciência da Informação no DF6_discentes_CI.

4.5.1.6.1 Fase 1 - Recorte dos dados da Ciência da Informação no DF3_trabalhos_CI

O recorte da área no DF3_trabalhos_BR foi realizado por meio do *script* apresentado no Quadro 95, que filtra os registros correspondentes à Ciência da Informação. O arquivo resultante contém 4.934 registros, todos com campos preenchidos.

Os dados foram revisados em planilha, identificando inconsistências no campo DS_RESUMO. Três registros apresentavam o valor "falta", e, após buscas na internet, apenas um resumo foi localizado. Os dois restantes permaneceram sem atualização. No Quadro 99, encontra-se o *script* empregado para extrair do DF3_trabalhos_BR apenas os registros associados à Ciência da Informação.

Quadro 99 – Script desenvolvido para recortar os dados da Ciência da Informação do DF3

```
import pandas as pd
# Caminho do arquivo original
df3_path = "/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF3/DF3.csv"
# Caminho do arquivo de saída
output_path = "/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF3/DF3_CI.csv"
# Carregar o arquivo CSV
print("Carregando DF3...")
df3 = pd.read_csv(df3_path, sep=";", encoding="utf-8")
# Filtrar os registros onde CD_AREA_CONHECIMENTO é 60700009
df3_filtered = df3[df3["CD_AREA_CONHECIMENTO"] == 60700009]
# Salvar o arquivo filtrado
df3_filtered.to_csv(output_path, sep=";", encoding="utf-8", index=False)
print(f"Arquivo filtrado salvo com sucesso: {output_path}")
```

Fonte: Dados de pesquisa (2024).

Foram também identificados resumos em inglês em trabalhos cujo idioma principal era o português. Após buscas na internet, os resumos em português foram localizados e atualizados. Além disso, caracteres especiais em alguns registros indicaram possíveis

problemas de encoding, que foram corrigidos. Os demais campos estão completos e prontos para as próximas etapas da análise.

4.5.1.6.2 Fase 2 - Recorte dos dados da Ciência da Informação no DF4_programas_CI

Para o recorte do DF4_programas_CI, foi desenvolvido um *script* para filtrar os registros da Ciência da Informação no DF4_programas_CI, conforme o Quadro 100.

Quadro 100 – Script desenvolvido para recortar os dados da Ciência da Informação do DF4_programas_CI

```
import pandas as pd
# Caminho do arquivo original
df4_path = "/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF4/DF4.csv"
# Caminho do arquivo de saída
output_path = "/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF4/DF4_CI.csv"
# Carregar o arquivo CSV
print("Carregando DF4...")
df4 = pd.read_csv(df4_path, sep=";", encoding="utf-8")
# Filtrar os registros onde CD_AREA_CONHECIMENTO é 60700009
df4_filtered = df4[df4["CD_AREA_CONHECIMENTO"] == 60700009]
# Salvar o arquivo filtrado
df4_filtered.to_csv(output_path, sep=";", encoding="utf-8", index=False)
print(f"Arquivo filtrado salvo com sucesso: {output_path}")
```

Fonte: Dados de pesquisa (2024).

O *script* cumpre seu propósito ao isolar os registros da Ciência da Informação no DF4_programas_CI, gerando um novo arquivo pronto para análise que contém 220 linhas. A verificação posterior em formato de planilha confirmou que todos os campos estavam completos e não houve necessidade de ajustes.

4.5.1.6.3 Fase 3 - Recorte dos dados da Ciência da Informação no DF5_docentes_CI

O recorte do DF5_docentes_CI foi realizado com base nos códigos dos programas identificados no DF3_trabalhos_BR. Para isso, foi aplicado um filtro que seleciona apenas os registros correspondentes à Ciência da Informação. O *script* apresentado no Quadro 101 realiza a leitura do arquivo DF5.csv, filtra as linhas cujo campo CD_PROGRAMA_IES corresponde à lista especificada e salva o resultado no arquivo DF5_CI.csv.

Quadro 101 – Script desenvolvido para recortar os dados da Ciência da Informação do DF5_docentes_CI5

```
import pandas as pd
# Caminho do arquivo original
df5_path = "/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF5/DF5.csv"
# Caminho do arquivo de saída
output_path = "/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF5/DF5_CI.csv"
# Lista dos valores a serem filtrados no campo CD_PROGRAMA_IES
valores_filtrar = [
```

```

"31003010056P0", "33006016003P4", "32053010003P1", "32001010028P2", "15001016158P5",
"28001010041P0", "31003010063P6", "22033017002P3", "33001014052P3", "33004110043P4",
"53001010018P0", "33002010195P5", "41001010052P0", "31021018018P8", "33301000001P0",
"31021018019P4", "27001016175P0", "24001015049P7", "41002016025P5", "23001011080P9",
"40002012048P2", "25001019077P3", "42001013176P0", "31001017138P0", "40002012035P8",
"32001010172P6", "22001018085P8", "26001012171P2", "30001013108P0", "33002010248P1",
"32053010006P0"
]
# Carregar o arquivo CSV
print("Carregando DF5...")
df5 = pd.read_csv(df5_path, sep=";", encoding="utf-8")
# Filtrar os registros onde CD_PROGRAMA_IES está na lista de valores
df5_filtered = df5[df5["CD_PROGRAMA_IES"].isin(valores_filtrar)]
# Salvar o arquivo filtrado
df5_filtered.to_csv(output_path, sep=";", encoding="utf-8", index=False)
print(f"Arquivo filtrado salvo com sucesso: {output_path}")

```

Fonte: Dados de pesquisa (2024).

Após a execução do *script*, o arquivo gerado contém 6.057 linhas. A verificação dos dados confirmou que o arquivo está completo, sem campos vazios, estando pronto para as análises subsequentes.

4.5.1.6.4 Fase 4 - Recorte dos dados da Ciência da Informação no DF6_discentes_CI

O DF6_discentes_CI foi recortado com base no CD_PROGRAMA_IES, utilizando o *script* apresentado no Quadro 102. O *script* lê o arquivo DF6.csv, aplica um filtro nos registros correspondentes à Ciência da Informação e salva o resultado no arquivo DF6_CI.csv.

Quadro 102 – Script desenvolvido para recortar os dados da Ciência da Informação do DF6

```

import pandas as pd
# Caminho do arquivo original
df6_path = "/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF6/DF6.csv"
# Caminho do arquivo de saída
output_path = "/content/drive/MyDrive/CAPES/datasets/dados_abertos/DF6/DF6_CI.csv"
# Lista dos valores a serem filtrados no campo CD_PROGRAMA_IES
valores_filtrar = [
    "31003010056P0", "33006016003P4", "32053010003P1", "32001010028P2", "15001016158P5",
    "28001010041P0", "31003010063P6", "22033017002P3", "33001014052P3", "33004110043P4",
    "53001010018P0", "33002010195P5", "41001010052P0", "31021018018P8", "33301000001P0",
    "31021018019P4", "27001016175P0", "24001015049P7", "41002016025P5", "23001011080P9",
    "40002012048P2", "25001019077P3", "42001013176P0", "31001017138P0", "40002012035P8",
    "32001010172P6", "22001018085P8", "26001012171P2", "30001013108P0", "33002010248P1",
    "32053010006P0"
]
# Carregar o arquivo CSV
print("Carregando DF6...")
df6 = pd.read_csv(df6_path, sep=";", encoding="utf-8", low_memory=False)
# Exibir todas as colunas para identificar o nome correto
print("Colunas encontradas:", df6.columns.tolist())
# Tentar encontrar a coluna correta, mesmo que tenha variações no nome
coluna_correta = None
for col in df6.columns:
    if "CD_PROGRAMA_IES" in col.strip():
        coluna_correta = col
        break
# Verificar se a coluna foi encontrada
if coluna_correta is None:
    raise KeyError("A coluna CD_PROGRAMA_IES não foi encontrada no DF6.csv")
# Filtrar os registros onde CD_PROGRAMA_IES está na lista de valores
df6_filtered = df6[df6[coluna_correta].astype(str).isin(valores_filtrar)]

```

```
# Salvar o arquivo filtrado
df6_filtered.to_csv(output_path, sep=";", encoding="utf-8", index=False)
print(f'Arquivo filtrado salvo com sucesso: {output_path}')
```

Fonte: Dados de pesquisa (2024).

Após a execução do *script*, o arquivo foi aberto em planilha para verificação. Durante a análise, foi identificada uma discrepância entre o DF3_trabalhos_CI e o DF6_discentes_CI. Enquanto o 'DF3_trabalhos_CI' apresenta 4.934 registros, o DF6_discentes_CI contém 4.926, resultando em uma diferença de 137 registros. A comparação dos arquivos revelou variações nos nomes dos discentes, incluindo diferenças na grafia, abreviações e erros de escrita. Além disso, o programa 32053010003P1 - FUMEC (2013 a 2017) estava cadastrado como multidisciplinar sob um código diferente, situação observada também em outros programas.

Outro fator analisado foi a situação dos discentes, com alguns registros marcados como "matriculado" em vez de "titulado". Para manter a coerência dos dados, esses casos foram ajustados. No entanto, dois discentes não foram encontrados nos arquivos originais de discentes. Como ambos possuem currículo Lattes e suas titulações registradas, optou-se por manter essa diferença de dois registros entre o DF3_trabalhos_CI e o DF6_discentes_CI.

4.5.1.7 Etapa 7 – Visualização dos dados

A etapa 7 é responsável pela visualização dos dados em forma de resultados. A visualização de dados em estudos bibliométricos tem a função de organizar e interpretar padrões e tendências em grandes volumes de informações, complementando a análise quantitativa e permitindo uma abordagem mais objetiva e acessível. Representações visuais, como gráficos, tabelas e mapas, facilitam a observação da evolução da produção científica, da distribuição geográfica dos pesquisadores e das variações institucionais e setoriais. Esse processo possibilita a identificação de lacunas e tendências emergentes, contribuindo para a formulação de novas perguntas de pesquisa.

No contexto deste estudo, que trata da produção acadêmica da pós-graduação brasileira entre 2004 e 2022, a visualização dos dados é realizada por meio de gráficos, tabelas e mapas, e possibilitará a análise da distribuição dos registros e a comparação entre diferentes períodos e áreas do conhecimento. Para isso, é aplicado um recorte específico à Ciência da Informação, permitindo avaliar sua evolução em relação ao cenário geral da pós-graduação. Esse recorte facilita a identificação de particularidades e tendências específicas da área, possibilitando uma análise direcionada. A segmentação dos dados também permite comparações entre mestrado e doutorado, tanto acadêmico quanto profissional, possibilitando

verificar diferenças nos padrões de titulação, tempo de conclusão e distribuição regional dos programas.

Além de favorecer a análise, a comunicação dos resultados se beneficia da organização visual dos dados, tornando as informações mais acessíveis a diferentes públicos e facilitando a disseminação do conhecimento científico. A representação gráfica contribui para uma abordagem exploratória detalhada, permitindo a identificação de distribuições, *outliers* e possíveis relações entre variáveis. Assim, a visualização dos dados não apenas sistematiza os achados, mas também fornece subsídios para uma interpretação mais precisa e uma avaliação mais abrangente do desenvolvimento da Ciência da Informação ao longo do período analisado.

4.5.2 Quadro síntese das etapas dos procedimentos metodológicos

Com o objetivo de facilitar a consulta e reforçar a sistematização das etapas desenvolvidas ao longo do tratamento e análise dos dados, apresenta-se o Quadro 103 que é uma síntese que reúne, de forma organizada, as sete etapas metodológicas implementadas nesta pesquisa.

Quadro 103 – Síntese das etapas metodológicas da pesquisa

Etapas	Descrição	Produtos Gerados
1. Coleta de Dados	Levantamento e download de 67 arquivos da CAPES (docentes, discentes, programas, teses e dissertações) no período de 2004 a 2022.	Conjuntos de dados brutos (DF1 e DF2)
2. Tratamento de Dados	Conversão e padronização dos dados, remoção de inconsistências, normalização terminológica e categorização.	Dados limpos e organizados em dataframes temáticos
3. Consolidação de Dados	Integração de DF1 e DF2 no DF3_trabalhos_BR, com estrutura comum de 24 campos.	DataFrame unificado (DF3_trabalhos_BR)
4. Criação do DataFrame	Estruturação e validação do DF3, incluindo geração de identificadores únicos e verificação de integridade.	DF3 finalizado com 1.148.250 registros
5. Análise Exploratória de Dados	Inspeção de valores nulos, outliers, padrões estatísticos, inconsistências e distribuição das variáveis.	Scripts de análise, gráficos e tabelas-resumo
6. Recorte de Dados	Filtragem dos dados para a área da Ciência da Informação com base em códigos da CAPES.	DF3_CI, DF4_CI, DF5_CI, DF6_CI
7. Visualização de Dados	Representação gráfica dos principais resultados com uso de Python (matplotlib/seaborn).	Quadros, tabelas, figuras e gráficos comparativos e painéis analíticos

Fonte: Dados da pesquisa (2025).

O Quadro 103 sintetiza as etapas metodológicas adotadas nesta pesquisa, oferecendo uma visão articulada do percurso seguido desde a coleta até a visualização dos dados. Essa sistematização não apenas favorece a compreensão do processo analítico empreendido, como

também forneceu subsídios diretos para a elaboração da proposta metodológica apresentada no capítulo 6, seção, 6.6, da Metodologia Integrada de Tratamento e Análise Bibliométrica de Dados Abertos (MITAB-DA), consolidando um modelo replicável para estudos bibliométricos baseados em dados abertos.

5 APRESENTAÇÃO DOS RESULTADOS

Neste capítulo são apresentados os resultados da pesquisa, obtidos a partir da análise dos dados da pós-graduação em Ciência da Informação no Brasil, com base nos dados abertos de teses e dissertações da CAPES. A apresentação dos resultados segue a estrutura das sete etapas metodológicas, a saber: coleta de dados, tratamento de dados, consolidação dos conjuntos de dados, criação do DataFrame Final (DF3), análise exploratória de dados (DF3_trabalhos_BR), recorte de dados e visualização dos dados. A exposição permite compreender as características gerais dos conjuntos de dados trabalhados, seguidas pela apresentação dos ajustes realizados para garantir a integridade e a consistência das informações. Essa abordagem possibilita uma visão detalhada do panorama da pós-graduação, evidenciando aspectos como produção acadêmica, distribuição geográfica, evolução temporal e outros indicadores relevantes, com foco na área da Ciência da Informação.

5.1 Resultados da etapa 1 Coleta de dados

Nesta etapa, foi explorado o ambiente de coleta de dados, especificamente o Catálogo de Teses e Dissertações (CTD) da CAPES, para identificar e selecionar os conjuntos de dados que comporiam a amostra da pesquisa. Retoma-se a informação de que a CAPES disponibiliza informações sobre a pós-graduação brasileira em seu porta¹⁴, organizadas em quatro grandes eixos: 1) Acessos ao Portal de Periódicos; 2) Bolsas e Auxílios; 3) Avaliação da Pós-Graduação *Stricto Sensu*; e 4) Servidores, Contratos, Dotação e Execução Orçamentária (Capes, 2022a).

Para este estudo, foi selecionado o eixo 3 – Avaliação da Pós-Graduação *Stricto Sensu* –, que contém 35 conjuntos de dados distribuídos em oito divisões temáticas. A Tabela 35 apresenta a organização desses eixos e a quantidade de conjuntos de dados disponíveis para a pós-graduação *stricto sensu*.

Tabela 35 – Divisões temáticas e conjuntos de dados sobre a avaliação da Pós-Graduação

Ordem	Divisão temática	Número de conjunto de dados
1	Detalhes da produção intelectual dos Programas da Pós-Graduação <i>strictu sensu</i> do Brasil	9
2	Catálogo de Teses e Dissertações	4

¹⁴ <https://dadosabertos.capes.gov.br/>

Ordem	Divisão temática	Número de conjunto de dados
3	Discentes da Pós-Graduação <i>Stricto Sensu</i> do Brasil	4
4	Docentes da Pós-Graduação <i>Stricto Sensu</i> do Brasil	4
5	Membros de Projetos dos Programas de Pós-Graduação <i>Stricto Sensu</i> do Brasil	4
6	Produção Intelectual dos Programas de Pós-Graduação <i>Stricto Sensu</i> do Brasil	4
7	Programas de Pós-Graduação <i>Stricto Sensu</i> do Brasil	3
8	Autor da Produção Intelectual de Programas de Pós-Graduação <i>Stricto Sensu</i> do Brasil	3

Fonte: Adaptado da CAPES (2022).

Dentre os conjuntos de dados expostos na Tabela 35, foram selecionados quatro conjuntos de dados para a pesquisa: 1) Catálogo de Teses e Dissertações – Brasil, 2) Docentes da Pós-Graduação *Stricto Sensu* no Brasil, 3) Discentes da Pós-Graduação *Stricto Sensu* no Brasil e 4) Programas de Pós-Graduação *Stricto Sensu* no Brasil.

O primeiro conjunto de dados, o Catálogo de Teses e Dissertações – Brasil, conforme o PDA da CAPES (2022), visa facilitar o acesso a informações sobre teses e dissertações defendidas nos Programas de Pós-Graduação do país e disponibilizar estatísticas sobre essa produção intelectual. Ele está dividido em quatro conjuntos de dados: de 1987 a 2012, de 2013 a 2016, de 2017 a 2020 e de 2021 a 2024. No último conjunto, até a data final da coleta de dados (março de 2025), havia apenas um arquivo referente ao ano de 2022. Esses catálogos disponibilizam arquivos em formato CSV e XLS para os dados e um arquivo em PDF contendo os metadados do banco de teses e dissertações (Capes, 2022a). Para este estudo, foram selecionados dezenove arquivos CSV, abrangendo os anos de 2004 a 2022. Os dados anteriores a 2004 (de 1987 a 2003) foram desconsiderados devido à ausência de informações sobre docentes e discentes nesses anos.

O segundo conjunto de dados refere-se aos Docentes da Pós-Graduação *Stricto Sensu* no Brasil. Esse conjunto contém informações sobre titulação acadêmica (graduação, mestrado, doutorado e notório saber), vínculo com a instituição (servidor público, regime celetista, entre outros) e vínculo com o programa (dedicação exclusiva, integral ou parcial) (Capes, 2022a). Os dados estão disponíveis nos formatos CSV, XLS, RDF e TTL, além de um arquivo PDF com os metadados. Foram selecionados dezenove arquivos CSV, abrangendo o período de 2004 a 2022. Informações sobre autores e produção intelectual, artística e técnica foram desconsideradas por não estarem no escopo deste estudo.

O terceiro conjunto de dados analisado corresponde aos Discentes da Pós-Graduação *Stricto Sensu* no Brasil, contendo dados sobre nacionalidade, ano de matrícula, nível de titulação, situação no programa, instituição de ensino vinculada, entre outras informações. Esses dados estão disponíveis nos formatos CSV e XLS, acompanhados de um arquivo PDF com os metadados dos discentes (Capes, 2022a). Foram selecionados dezenove arquivos CSV, abrangendo os anos de 2004 a 2022, que é o período estabelecido como recorte.

O quarto e último conjunto de dados analisado corresponde aos Programas de Pós-Graduação *Stricto Sensu* no Brasil. Esse catálogo contém informações sobre os programas, incluindo código, nome, ano de início, siglas e nomes das instituições vinculadas, área de avaliação e modalidade (acadêmica ou profissional). Os dados estão disponíveis nos formatos CSV e XLS, além de um arquivo PDF com os metadados (Capes, 2022a). Diferente dos demais conjuntos, esse catálogo abrange apenas o período de 2013 a 2022, resultando na seleção de dez arquivos CSV.

No total, foram obtidos 67 arquivos CSV, distribuídos entre os quatro conjuntos de dados: 19 do Catálogo de Teses e Dissertações, 19 de docentes, 19 de discentes e 10 dos programas de pós-graduação. Além desses, foram baixados quatro arquivos de metadados, um para cada conjunto, contendo descrições detalhadas sobre a estrutura e o conteúdo dos dados, essenciais para sua correta interpretação e processamento.

5.2 Resultados da etapa 2 Tratamento dos dados

Nesta etapa, o tratamento dos dados envolveu a organização, padronização e estruturação dos conjuntos coletados para garantir sua integridade e viabilidade para análise. Os procedimentos consistiram na conversão dos arquivos para o formato CSV, permitindo a compatibilização dos dados entre os diferentes conjuntos. Os arquivos foram importados para o ambiente do Google Colab¹⁵ PRO, onde foram processados utilizando a linguagem Python e bibliotecas especializadas, como pandas (manipulação de dados), os (gestão de diretórios e arquivos), unidecode (padronização de caracteres) e time (controle de tempo de execução).

Durante a triagem, os dados foram revisados sendo identificadas inconsistências nos registros, incluindo valores ausentes, duplicidades e variações na grafia de nomes e siglas. Para mitigar esses problemas, aplicaram-se técnicas de normalização, como a remoção de caracteres especiais, conversão de textos para caixa alta ou baixa e a substituição de valores nulos por identificadores padrão. Foram aplicadas técnicas de categorização e codificação,

¹⁵ <https://colab.research.google.com/>

permitindo a organização das informações conforme as variáveis definidas. Também foram realizadas filtragens, garantindo que apenas registros completos e relevantes fossem mantidos.

Ao final do tratamento, os dados foram organizados em oito dataframes, sendo quatro referentes ao panorama geral da pós-graduação *stricto sensu* no Brasil e quatro específicos da Ciência da Informação. Essa estrutura possibilitou uma abordagem comparativa, facilitando a análise das tendências e características da área em relação ao sistema nacional de pós-graduação.

5.3 Resultados da etapa 3 de Consolidação dos conjuntos de dados

Nesta etapa 3, a consolidação dos dados foi conduzida em cinco fases, garantindo a integração dos diferentes conjuntos e a estruturação dos DataFrames para as análises posteriores.

Na primeira fase, os DataFrames DF1 e DF2 foram estruturados a partir da identificação e mapeamento das variáveis comuns e exclusivas, validando a correspondência entre os conjuntos. Esse processo garantiu a definição do esquema do DF3, que unifica as informações essenciais para a pesquisa.

A segunda fase consistiu na agregação dos dados de 2004 a 2012 em um único conjunto, padronizando nomenclaturas, eliminando redundâncias e incorporando campos necessários para compatibilização com os conjuntos posteriores. Foram ajustadas as variáveis para que os dados desse período se alinhassem estruturalmente aos registros mais recentes.

Na terceira fase, os dados agregados de 2004 a 2012 foram incorporados ao DF1, com tratamento de valores ausentes, ajustes na estrutura dos campos e padronização da codificação para UTF-8. Esse processo assegurou a consistência dos registros e permitiu a uniformização das informações ao longo dos diferentes anos analisados.

A quarta fase incluiu o acoplamento dos dados de 2013 a 2022 no DF2, seguindo os mesmos procedimentos aplicados ao DF1. Foram identificadas discrepâncias entre as variáveis ao longo do tempo, exigindo ajustes na nomenclatura e a exclusão de campos de baixa representatividade. Essa etapa permitiu alinhar a estrutura do DF2 aos padrões definidos para o DF3.

Na quinta fase, os DF1 e DF2 foram normalizados para uniformizar as diferenças estruturais, considerando a evolução das variáveis ao longo dos anos. Foram definidos os campos a serem mantidos, ajustados ou descartados, garantindo uma integração consistente dos conjuntos de dados.

Como resultado, a consolidação permitiu a construção de DataFrames estruturados e compatíveis para a formação do DataFrame consolidado, apresentado na próxima etapa. O processo seguiu critérios metodológicos, assegurando a coerência das informações e a preparação dos dados para as análises da pós-graduação *stricto sensu* no Brasil no período de 2004 a 2022.

5.4 Resultados da etapa 4 de Criação do DataFrame (DF3_trabalhos_BR)

A etapa 4 consistiu na criação do DataFrame DF3_trabalhos_BR, consolidando os dados dos DataFrames DF1 e DF2 em um único DataFrame estruturado. O processo foi conduzido, inicialmente, pela estruturação dos dados, e, em seguida, pela consolidação, com a integração dos dados. Desse modo, primeiramente foi realizada a definição da composição do DF3_trabalhos_BR, organizando suas variáveis para garantir coerência com os conjuntos de dados originais. Isso incluiu a padronização de nomenclaturas, a adaptação de identificadores únicos e a normalização das variáveis. Foram eliminadas colunas redundantes e inseridas novas variáveis para manter a compatibilidade com o modelo de dados consolidado.

Na integração dos dados foi aplicada uma metodologia de migração sequencial. Inicialmente, os registros do DF1 foram transferidos para o DF3_trabalhos_BR, seguidos pelos do DF2. Esse processo foi conduzido por meio de um *script* em Python, utilizando a biblioteca Pandas para renomear colunas, gerar identificadores únicos e converter valores categóricos em códigos numéricos. Além disso, foi implementado um processo de verificação de integridade para avaliar a estrutura do DataFrame e identificar possíveis inconsistências nos dados. A incorporação do DF2 ao DF3_trabalhos_BR seguiu uma abordagem semelhante, mas com menor necessidade de transformações, pois sua estrutura já estava alinhada ao formato final. A junção foi realizada por concatenação dos registros, garantindo a preservação dos campos definidos anteriormente.

Após a integração, foram conduzidas análises para verificar a completude dos dados e identificar possíveis lacunas, resultando em um DataFrame final com 1.148.250 registros distribuídos em 24 colunas. Essa consolidação permitiu a estruturação dos dados para as etapas seguintes, viabilizando análises exploratórias e comparações entre diferentes dimensões da pós-graduação *stricto sensu* no Brasil.

5.5 Resultados da etapa 5 Análise exploratória dos dados do DF3

A etapa 5 da pesquisa envolveu a análise exploratória dos dados do DataFrame DF3, compreendendo cinco fases, a saber: Fase 1: Entendimento Inicial dos Dados, Fase 2: Verificação de Valores Ausentes, Fase 3: Entendimento profundo dos dados, Fase 4: Preparação dos dados e Fase 5: Análise exploratória profunda dos dados. A condução dessas fases permitiu avaliar a estrutura, identificar inconsistências e preparar os dados para as análises subsequentes.

Na primeira fase, foi realizado o entendimento inicial dos dados, verificando-se a estrutura, o tamanho do conjunto de dados e as variáveis presentes. O DF3 contém 1.232.870 registros distribuídos em 24 colunas, das quais três são do tipo inteiro, três do tipo *float* e 18 do tipo objeto. Um script em Python foi utilizado para carregar os dados, gerar estatísticas descritivas e contar a quantidade de valores únicos por coluna. Para facilitar a manipulação, foi adicionada ao DF3 uma coluna de índice único, permitindo a localização e recuperação de registros específicos. Durante essa análise inicial, identificou-se a presença de valores ausentes em algumas colunas, como NM_GRANDE_AREA_CONHECIMENTO e NM_GRAU_ACADEMICO.

A segunda fase consistiu na verificação e tratamento de valores ausentes. Um comando em Python foi aplicado para detectar colunas com registros incompletos. Os campos CD_GRANDE_AREA_CONHECIMENTO e NM_GRANDE_AREA_CONHECIMENTO apresentaram inconsistências na correspondência entre códigos e nomes das áreas. Outras colunas, como ID_PESSOA_DISCENTE e NM_DISCENTE, também continham dados faltantes. Para lidar com essas lacunas, estratégias como imputação de valores e exclusão de registros foram consideradas conforme a relevância da variável para as análises subsequentes. A relação entre as colunas foi examinada para entender a origem dos valores ausentes e definir abordagens adequadas para cada caso.

Na terceira fase, foi conduzida uma análise detalhada das variáveis, incluindo estatísticas descritivas, distribuição dos dados e identificação de padrões. *Scripts* foram utilizados para computar a contagem de frequência dos valores em determinadas colunas, como ANO_BASE, permitindo avaliar a evolução temporal das produções acadêmicas. Uma função da biblioteca Pandas foi empregada para mensurar a diversidade de valores em cada campo, identificando discrepâncias e sugerindo possíveis ajustes. Durante essa análise, verificou-se que algumas variáveis apresentavam altos níveis de cardinalidade, enquanto outras exibiam baixa variação, indicando possíveis ajustes. Além disso, a inspeção dos nomes

das áreas do conhecimento evidenciou discrepâncias entre os códigos e suas denominações, o que exigiu a criação de um *script* para verificar e corrigir inconsistências.

A quarta fase focou na organização e padronização dos dados para garantir a integridade das informações, abrangendo três passos distintos, cada um voltado para a organização e padronização dos conjuntos de dados: entendimento e ajustes para conjunto de dados dos programas, dos docentes e dos discentes. No Passo 1, os dados dos programas foram analisados e padronizados, assegurando que os identificadores estivessem corretamente vinculados às demais informações do estudo. Inicialmente, foi realizada a verificação da sua estrutura, identificando possíveis inconsistências ou registros duplicados. Em seguida, os ajustes incluíram a padronização dos identificadores dos programas, assegurando que cada um estivesse corretamente vinculado às demais informações do estudo. No Passo 2, os dados dos docentes passaram por normalização das informações, garantindo que os atributos, como vínculos institucionais e áreas de atuação, estivessem estruturados corretamente. No Passo 3, a abordagem foi semelhante, e os dados dos discentes foram verificados e uniformizados, com ajustes em valores ausentes e eliminação de inconsistências que poderiam afetar a análise comparativa entre os anos. Como resultado dessa fase, os três conjuntos de dados passaram a apresentar maior coerência e integridade.

Na quinta e última fase, foi realizada uma análise exploratória aprofundada do DF3, utilizando técnicas estatísticas para compreender padrões e relações entre variáveis. A distribuição dos dados foi visualizada por meio de gráficos e tabelas geradas com Python. Foram analisadas correlações entre variáveis para identificar tendências nos programas de pós-graduação ao longo dos anos. Discrepâncias entre categorias e a distribuição dos títulos acadêmicos foram investigadas, proporcionando uma visão do comportamento dos dados. Além disso, a detecção de valores atípicos e a avaliação da consistência dos registros contribuíram para a melhoria da qualidade dos dados.

Como resultado da etapa 5, os dados foram refinados, organizados e padronizados, garantindo maior confiabilidade para a realização das análises comparativas e estatísticas nas etapas seguintes da pesquisa.

5.6 Resultados da etapa 6 – Recorte de dados

A Etapa 6 consistiu no recorte dos dados da Ciência da Informação a partir dos conjuntos de dados processados nas etapas anteriores. O objetivo foi garantir que apenas os registros pertinentes a essa área fossem selecionados para análise. Como a estrutura dos

DataFrames variava e algumas variáveis estavam ausentes em determinados conjuntos, a fusão completa dos dados não foi viável. Dessa forma, optou-se por manter os DataFrames separados, resultando nos seguintes arquivos específicos para a área:

1. DF3_trabalhos_CI – contendo os registros de teses e dissertações da Ciência da Informação;
2. DF4_programas_CI – reunindo os programas de pós-graduação da área;
3. DF5_docentes_CI – com os dados dos docentes vinculados à pós-graduação em Ciência da Informação;
4. DF6_discentes_CI – incluindo os registros dos discentes matriculados nos programas da área.

O recorte foi realizado com base no campo da área de conhecimento, garantindo a seleção precisa dos registros relacionados à Ciência da Informação. Cada DataFrame foi processado individualmente, aplicando-se filtros específicos para manter apenas as informações relevantes.

A segmentação dos dados possibilita uma análise mais detalhada da pós-graduação na área, permitindo a identificação de padrões e tendências na produção acadêmica. A organização e o refinamento dos registros foram conduzidos de maneira a garantir que a visualização dos dados, apresentada na etapa sete, seja clara e representativa, facilitando a interpretação e a extração de informações sobre a evolução da área ao longo do período analisado. Essa abordagem assegura que os dados estejam estruturados para análises comparativas, permitindo um posicionamento da Ciência da Informação em relação ao panorama geral da pós-graduação no Brasil.

5.7 Resultado da etapa 7 de Visualização de dados

Nesta etapa, os resultados são apresentados em duas partes distintas. Primeiramente, são demonstrados os resultados referentes à pós-graduação brasileira, de modo geral, utilizando os quatro DataFrames processados:

1. DF3_trabalhos_BR – registros de teses e dissertações defendidas no Brasil (5.7.1);
2. DF4_programas_BR – informações sobre os programas de pós-graduação no Brasil (5.7.2);
3. DF5_docentes_BR – dados sobre os docentes vinculados aos programas de pós-graduação no Brasil (5.7.3);

4. DF6_discentes_BR – registros dos discentes matriculados e titulados nos programas de pós-graduação no Brasil (5.7.4).

Em seguida, é apresentado o recorte específico da Ciência da Informação, com base em outros quatro DataFrames:

1. DF3_trabalhos_CI – teses e dissertações da Ciência da Informação (5.7.5);
2. DF4_programas_CI – programas de pós-graduação da área da Ciência da Informação (5.7.6);
3. DF5_docentes_CI – docentes vinculados aos programas de pós-graduação em Ciência da Informação (5.7.7);
4. DF6_discentes_CI – discentes vinculados aos programas de pós-graduação em Ciência da Informação (5.7.8).

Essa organização permite tanto uma visão geral sobre a pós-graduação no Brasil e da Ciência da Informação entre 2004 e 2022, de modo isolado, quanto uma análise comparativa entre o cenário geral da pós-graduação brasileira e a realidade específica da Ciência da Informação no Brasil.

A segmentação dos conjuntos de dados facilita a identificação de padrões e tendências ao longo do período analisado, permitindo a observação de características próprias da área de Ciência da Informação e sua relação com o cenário mais abrangente da pós-graduação brasileira.

A visualização dos dados estruturados permite compreender o posicionamento da Ciência da Informação dentro do panorama da pós-graduação brasileira, subsidiando a análise de sua evolução, desafios e perspectivas.

5.7.1 Resultados do conjunto de dados DF3_trabalhos_BR

Os dados do DF3_trabalhos_BR totalizam 1.232.870 registros das teses e dissertações da pós-graduação *stricto sensu* no Brasil entre 2004 e 2022. As variáveis analisadas incluem o ano base, a grande área do conhecimento, a área do conhecimento, a área de avaliação, o nome do programa, a entidade de ensino, a região, a unidade federativa (UF), o título do trabalho, o resumo e o idioma. Esses campos possibilitam análises isoladas e cruzadas sobre a produção acadêmica no âmbito da pós-graduação brasileira ao longo do período trabalhado.

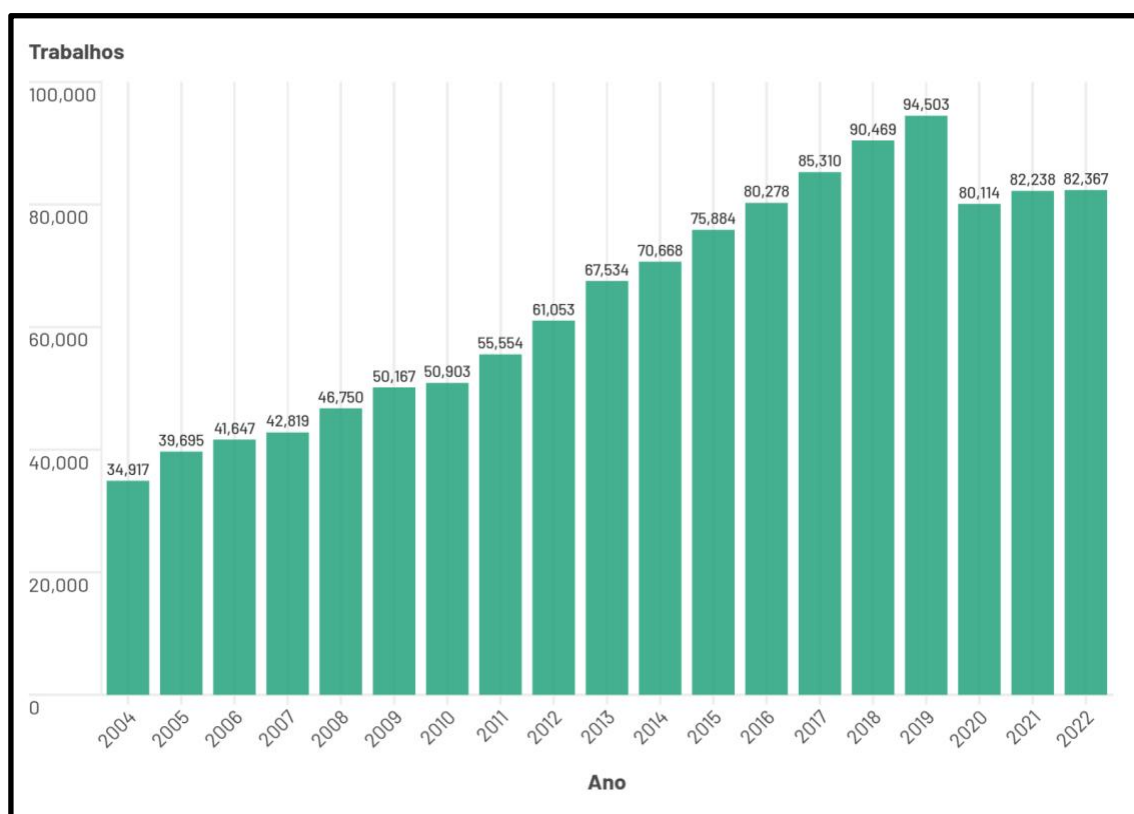
Observa-se um crescimento contínuo na produção acadêmica ao longo dos anos, com variações percentuais anuais que refletem diferentes contextos institucionais e sociais. Entre

2004 e 2019, o número passou de 34.917 para 94.503, representando um crescimento de aproximadamente 170,7% nesse período. Os anos de 2005, 2012 e 2013 apresentaram os maiores acréscimos percentuais, com variações de 13,68%, 9,89% e 10,61%, respectivamente.

Em 2020, houve uma redução de 15,23% no número de trabalhos, totalizando 80.114 registros, reflexo do impacto da pandemia da COVID-19. Nos anos seguintes, observou-se uma recuperação parcial, com um aumento de 2,65% em 2021 e uma estabilidade em 2022, com variação de apenas 0,16% em relação ao ano anterior.

Essa trajetória pode ser visualizada no Gráfico 8, que evidencia a retração na produção acadêmica em 2020 e a retomada gradual nos anos subsequentes.

Gráfico 8 – Número de trabalhos defendidos na pós-graduação no período de 2004 a 2022



Fonte: Dados de pesquisa (2025).

A análise do gráfico 8 indica que a pós-graduação brasileira passou por um período de crescimento expressivo, seguido por uma desaceleração em 2020, provavelmente devido à pandemia, com sinais de leve recuperação nos anos seguintes, mas mantendo números ainda sem atingir o ano de 2017. Para uma compreensão mais profunda, seria relevante, no futuro, investigar fatores institucionais, econômicos e políticos que eventualmente tenha impactado essas variações.

No Gráfico 9, observa-se a distribuição dos trabalhos defendidos na pós-graduação brasileira entre 2004 e 2022, classificados por grandes áreas do conhecimento, conforme propostas pela CAPES.

Gráfico 9 – Distribuição das Grandes Áreas do Conhecimento em percentual de 2004 a 2022



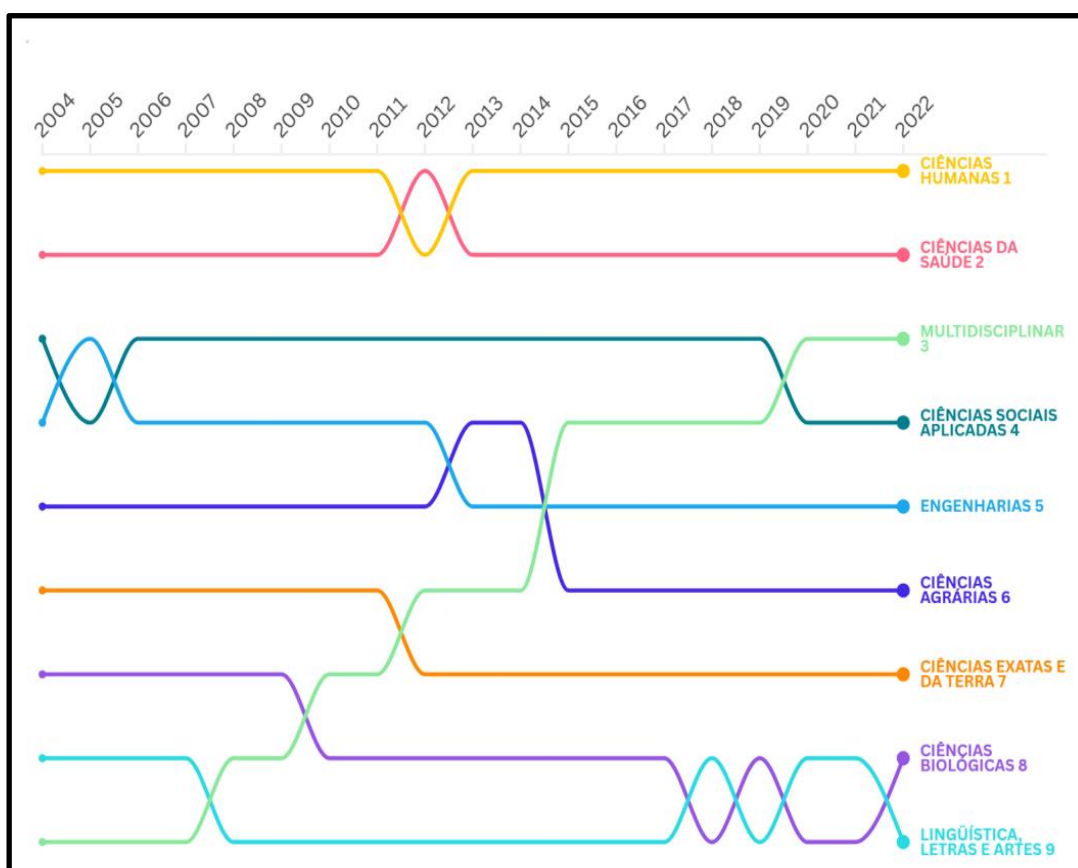
Fonte: Dados de pesquisa (2025).

A área com o maior volume de produções acadêmicas é Ciências Humanas, com 203.870 registros (16,54%). Em seguida, destacam-se Ciências da Saúde, com 192.624 trabalhos (15,62%) e Ciências Sociais Aplicadas, com 172.924 registros (14,03%). Juntas, essas três áreas concentram 46,19% de toda a produção acadêmica registrada no período analisado.

As Engenharias somam 137.891 trabalhos (11,18%), acompanhadas pela área Multidisciplinar, com 135.352 trabalho, 11% da produção total, e pelas Ciências Agrárias, com 122.472 trabalhos (9,93%). As Ciências Exatas e da Terra respondem por 8,57%, com 105.648 registros, enquanto as Ciências Biológicas somam 85.756 trabalhos (6,96%). Por fim, Linguística, Letras e Artes apresenta a menor participação relativa, com 76.333 registros (6,1%). A distribuição evidencia uma predominância das áreas das humanidades e ciências sociais na produção acadêmica da pós-graduação *stricto sensu* no Brasil, enquanto os campos das ciências exatas, biológicas e linguísticas têm participação proporcionalmente menor no conjunto analisado.

No Gráfico 10, observa-se a variação na posição das grandes áreas do conhecimento em termos de número de produção acadêmica ao longo do período analisado. Em 2004, Ciências Humanas liderava, seguida por Ciências da Saúde e Ciências Sociais Aplicadas. Até 2022, essa configuração se alterou, com Ciências da Saúde assumindo a liderança, Ciências Humanas em segundo lugar, e Ciências Sociais Aplicadas mantendo-se entre as primeiras posições. A área Multidisciplinar teve destaque, saltando da última colocação em 2004 para o quarto lugar em 2022, refletindo sua expressiva expansão na produção acadêmica. Por outro lado, áreas como Engenharias, Ciências Biológicas e Ciências Agrárias perderam posições, indicando relativa estagnação ou retração.

Gráfico 10 – Ranking das Áreas do Conhecimento por ano de 2004 a 2022

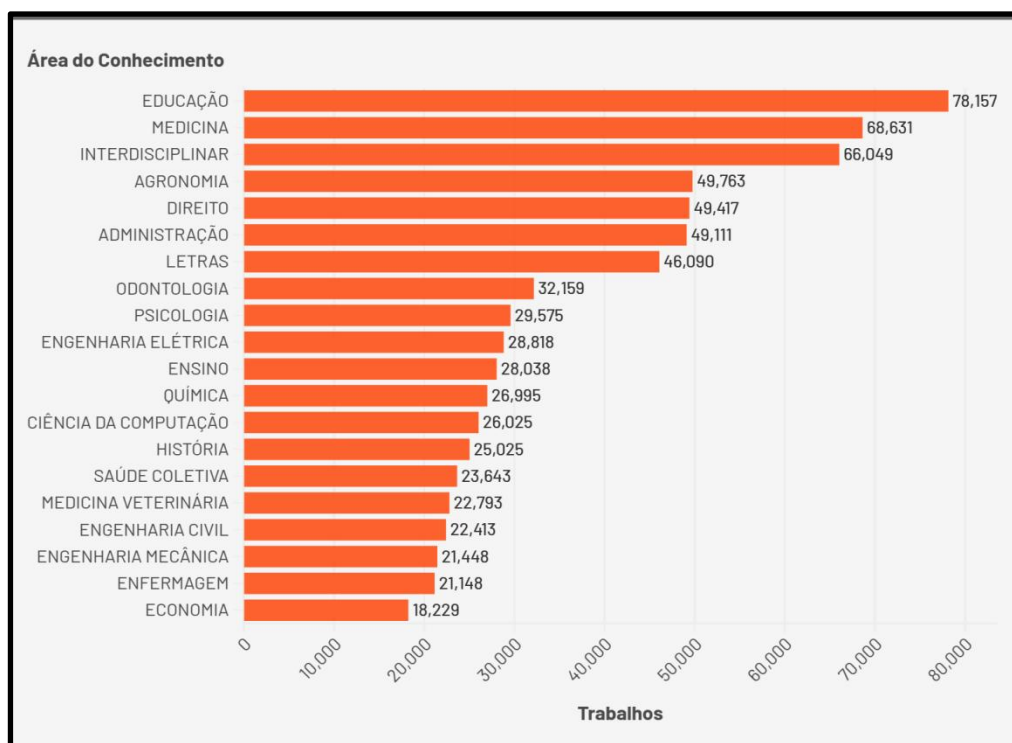


Fonte: Dados de pesquisa (2025).

O crescimento da produção acadêmica foi mais acentuado em áreas como: Multidisciplinar, que passou de 1.580 trabalhos em 2004 para 12.536 em 2022; Ciências da Saúde, de 5.327 para 13.093; Ciências Sociais Aplicadas, de 5.303 para 12.521. Em contraste, algumas áreas registraram queda no volume de produções: Engenharias, que caiu de um pico de 9.833 em 2019 para 7.457 em 2022; Ciências Biológicas, de 5.968 em 2019 para 5.110 em 2022; Ciências Agrárias, de 8.631 (2019) para 6.872 (2022).

No Gráfico 11, são apresentadas as 20 áreas do conhecimento com maior número de trabalhos defendidos entre 2004 e 2022, permitindo uma análise comparativa entre as áreas e suas respectivas grandes áreas do conhecimento.

Gráfico 11 – 20 áreas do conhecimento com maior número de trabalhos de 2004 a 2022



Fonte: Dados de pesquisa (2025).

A área com maior volume de trabalhos defendidos no período de 2004 a 2022 é Educação, vinculada à grande área das Ciências Humanas, com 78.157 registros. Em seguida, Medicina, da área de Ciências da Saúde, contabiliza 68.631 trabalhos, e a área Interdisciplinar, pertencente à grande área Multidisciplinar, soma 66.049 registros. Agronomia, vinculada às Ciências Agrárias, totaliza 49.763 trabalhos, enquanto Direito e Administração, ambas da grande área Ciências Sociais Aplicadas, apresentam 49.417 e 49.111 registros, respectivamente. Letras, pertencente à grande área Linguística, Letras e Artes, contabiliza 46.090 trabalhos.

No conjunto das Ciências da Saúde, além de Medicina, destacam-se ainda Odontologia com 32.159, Saúde Coletiva com 23.643 e Enfermagem com 21.148 trabalhos. As Engenharias também figuram entre as áreas com maior número de produções acadêmicas, com destaque para Engenharia Elétrica com 28.818, Engenharia Civil com 22.413 e Engenharia Mecânica com 21.448 registros. Nas Ciências Exatas e da Terra, a Química registra 26.995 trabalhos e a Ciência da Computação, 26.025. Entre as Ciências Humanas,

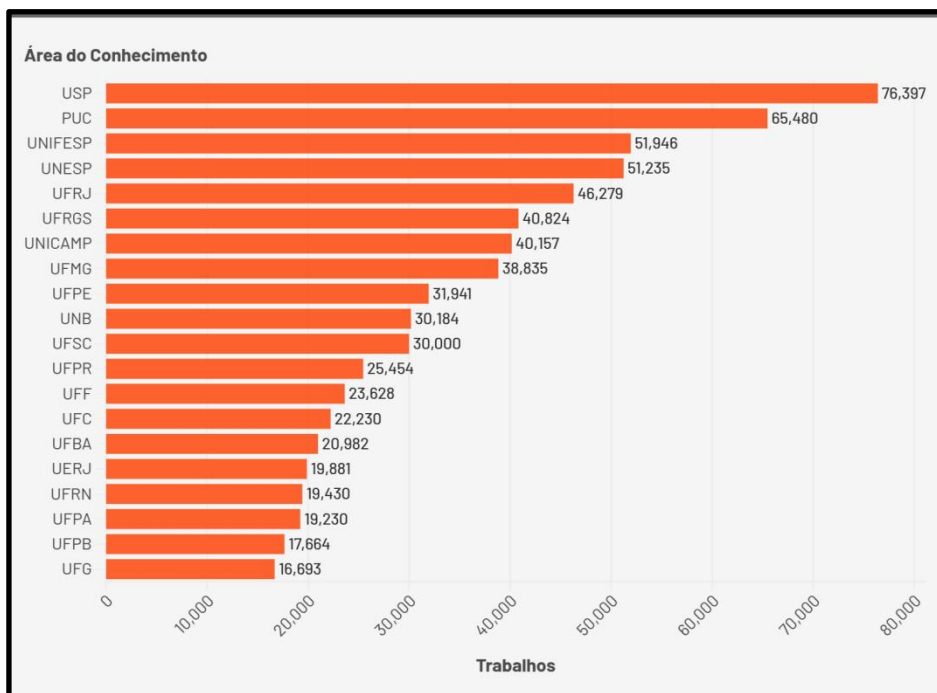
destacam-se ainda Psicologia com 29.575 e História com 25.025 registros. Já Medicina Veterinária, das Ciências Agrárias, apresenta 22.793 registros, e Economia, das Ciências Sociais Aplicadas, soma 18.229.

A distribuição evidencia que a maior concentração de produções acadêmicas está nas grandes áreas das Ciências Humanas, Ciências da Saúde, Ciências Sociais Aplicadas e Engenharias, com participação relativamente menor das áreas ligadas às Ciências Exatas, Biológicas e Agrárias.

Cabe observar que o campo referente às áreas do conhecimento apresenta um total de 248 valores distintos, o que dificulta a representação gráfica completa. Além disso, a nomenclatura adotada nos registros apresenta sobreposições e variações, como é o caso de Administração, Administração de Empresas e Administração Pública, ou ainda de Matemática e Matemática Aplicada, o que pode comprometer a clareza na interpretação direta dos dados e reforça a necessidade de padronização terminológica.

A análise das Instituições de Ensino Superior (IES) no período de 2004 a 2022 indica uma concentração da produção acadêmica em um grupo restrito de universidades, embora o conjunto de dados abranja 491 instituições distintas. Como mostra o Gráfico 12, a Universidade de São Paulo (USP) lidera em número de trabalhos defendidos, com 76.397 registros, seguida pela Pontifícia Universidade Católica (PUC), com 65.480. Entre as universidades federais, destacam-se a Universidade Federal de São Paulo (UNIFESP), com 51.946 registros, e a Universidade Estadual Paulista Júlio de Mesquita Filho (UNESP), com 51.235. A Universidade Federal do Rio de Janeiro (UFRJ-IBICT) aparece na quinta posição, com 46.279 trabalhos. Esses dados indicam que a maior parte da produção científica nacional está concentrada em universidades públicas e instituições tradicionais com forte vocação para ensino e pesquisa.

Gráfico 12 – 20 IES com o maior número de trabalhos de 2004 a 2022

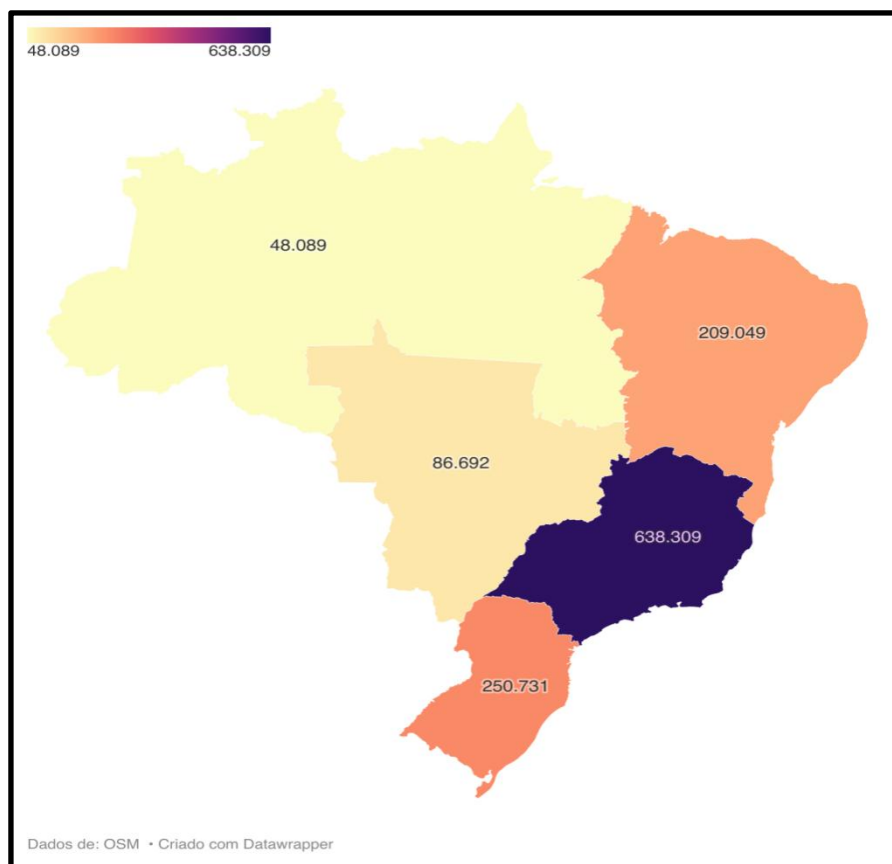


Fonte: Dados de pesquisa (2025).

A predominância das universidades federais no ranking das 20 instituições com maior número de trabalhos defendidos reforça a importância dessas instituições para a pós-graduação no Brasil. A Universidade Federal do Rio Grande do Sul (UFRGS), Universidade Estadual de Campinas (UNICAMP) e Universidade Federal de Minas Gerais (UFMG) registram, respectivamente, 40.824, 40.157 e 38.835 trabalhos. A Universidade Federal de Pernambuco (UFPE), Universidade de Brasília (UNB) e Universidade Federal de Santa Catarina (UFSC) também apresentam produção acadêmica expressiva, com valores entre 30.000 e 31.941 registros. Na segunda metade da lista de posições, a presença de universidades como a Universidade Federal do Paraná (UFPR), Universidade Federal Fluminense (UFF) e Universidade Federal do Ceará (UFC) demonstra a distribuição da produção científica por diferentes regiões do país. A concentração dos trabalhos em universidades públicas, especialmente federais e estaduais, confirma o papel estruturante dessas instituições no sistema nacional de pós-graduação e na produção científica no Brasil.

O Gráfico 13 mostra a distribuição dos trabalhos defendidos entre 2004 e 2022, evidenciando uma concentração da produção acadêmica nas regiões mais desenvolvidas economicamente.

Gráfico 13 – Distribuição dos trabalhos defendidos por regiões do país, no período 2004-2022



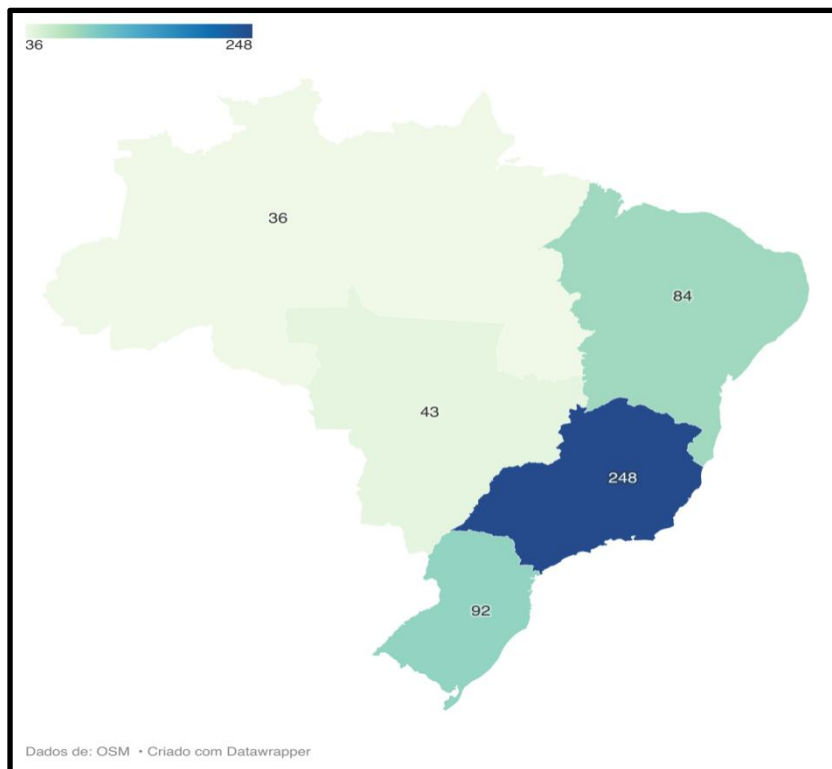
Fonte: Dados de pesquisa (2025).

Observa-se que a Região Sudeste concentra o maior volume de trabalhos defendidos, totalizando 638.309 registros (51,77%) da produção acadêmica nacional. Em seguida, a Região Sul apresenta 250.731 trabalhos (20,34%). Juntas, essas duas regiões respondem por 72,11% do total de produções acadêmicas registradas no período analisado. As demais regiões apresentam participação proporcionalmente menor. A Região Nordeste contabiliza 209.049 registros (16,96%), enquanto a Região Centro-Oeste soma 86.692 trabalhos (7,03%). A Região Norte apresenta o menor volume, com 48.089 registros, correspondendo a 3,90% do total. Esses dados evidenciam um desequilíbrio regional na pós-graduação brasileira, refletindo possíveis disparidades no sistema de ensino superior, como a distribuição desigual de instituições, a oferta de cursos em programas de pós-graduação e o acesso diferenciado a recursos e financiamento para pesquisa.

Complementando a análise da distribuição da produção acadêmica por região, o Gráfico 14 apresenta o número de instituições de ensino superior ativas entre 2004 e 2022. Observa-se que a Região Sudeste concentra o maior número de instituições (248), acompanhando sua posição de destaque em volume de trabalhos defendidos. Em seguida, vem a Região Sul, com 92 instituições, e o Nordeste, com 84, que também apresentam níveis

expressivos de produção acadêmica. Por sua vez, com menores quantidades, vem o Centro-Oeste, que contabiliza 43 instituições, e a Região Norte, que registra 36.

Gráfico 14 – Distribuição das IES por região, no período 2004-2022



Fonte: Dados de pesquisa (2025).

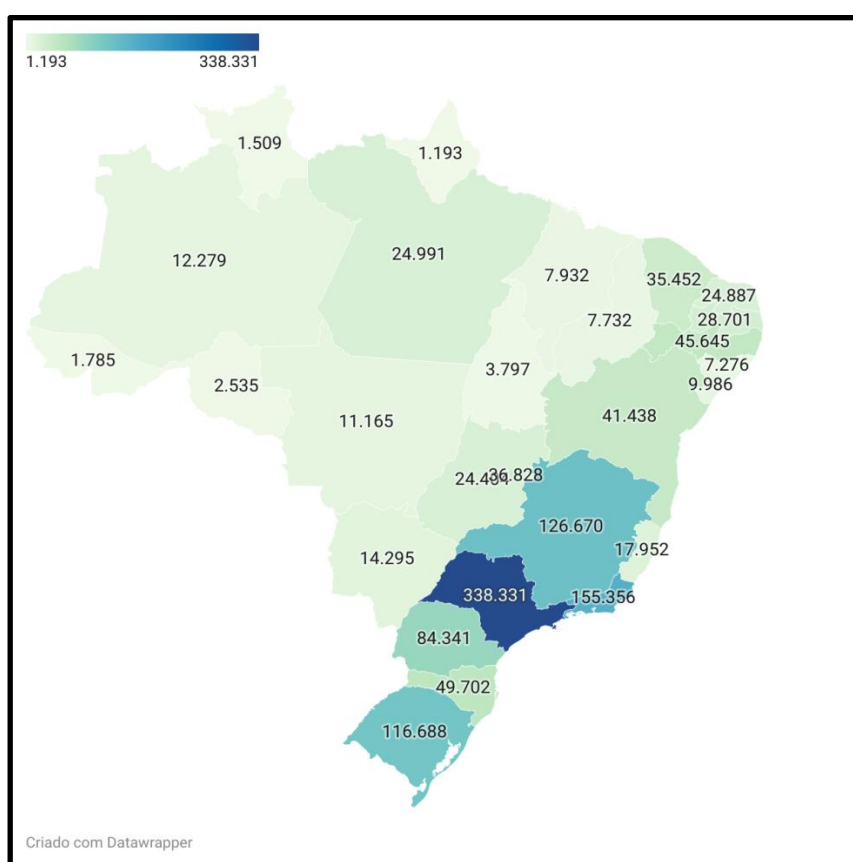
A comparação entre o número de instituições e o volume de produção mostra que a distribuição dos programas de pós-graduação não é proporcional em todas as regiões. No Sudeste, por exemplo, 49,3% das instituições respondem por 51,77% da produção nacional, indicando maior densidade de programas e maior média de trabalhos defendidos por instituição. No Norte, a discrepância é inversa: 7,15% das instituições respondem por apenas 3,90% da produção, o que sugere menor produtividade média por instituição. Nas regiões Nordeste e Centro-Oeste, observa-se uma correspondência mais equilibrada entre número de instituições e volume de produção, apontando para uma distribuição proporcional dos programas e de sua produtividade. Essa análise reforça a presença de desigualdades regionais na estrutura da pós-graduação, que podem estar associadas a fatores como financiamento, infraestrutura de pesquisa, concentração de docentes e maturidade institucional.

Cabe destacar que o conjunto de dados considera 491 instituições distintas, com base em registros únicos por nome. No entanto, o somatório das instituições por região totaliza 503, refletindo uma diferença de 12 registros. Essa discrepância ocorre porque algumas instituições mantêm unidades em diferentes regiões do país, sendo contabilizadas mais de uma vez.

conforme sua atuação geográfica. É o caso, por exemplo, da Fundação Oswaldo Cruz (FIOCRUZ), da Pontifícia Universidade Católica (PUC) e da Fundação Getúlio Vargas (FGV), que possuem estabelecimentos em múltiplos estados e contribuem para a produção acadêmica em mais de uma localidade.

Os dados regionais apresentados no Gráfico 15 podem ser correlacionados com os dados da distribuição dos trabalhos defendidos por unidade da federação, mostrados no Gráfico 14, permitindo a identificação dos estados que concentram a maior parte da produção acadêmica em cada região.

Gráfico 15 – Quantidade de trabalhos defendidos por UF, no período 2004-2022



Fonte: Dados de pesquisa (2025)

A análise do Gráfico 15 confirma que a Região Sudeste, responsável por 638.309 trabalhos defendidos (51,77% do total nacional), concentra sua produção majoritariamente em São Paulo (SP), com 338.331 registros (27,44%), seguido por Rio de Janeiro (RJ), com 155.356 (12,60%), e Minas Gerais (MG), com 126.670 (10,27%). O Espírito Santo (ES), embora com menor representatividade, contabiliza 17.952 trabalhos (1,46%). Esses dados refletem a forte presença de universidades e centros de pesquisa consolidados na região.

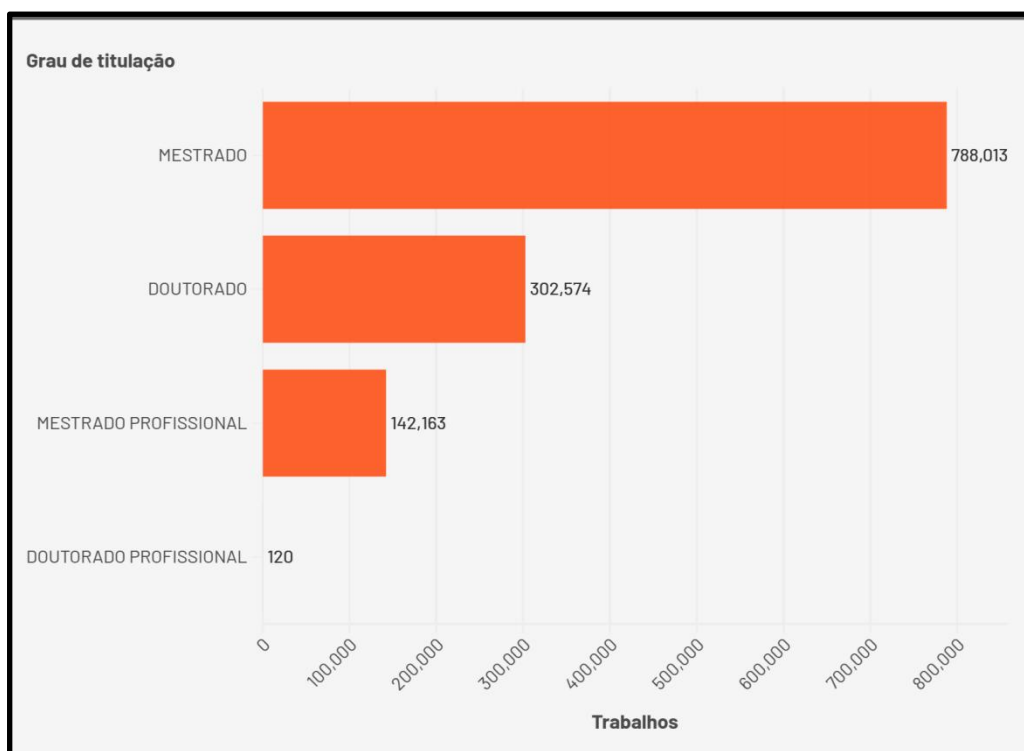
No Sul, que reúne 250.731 trabalhos (20,34%), o Rio Grande do Sul (RS) lidera com 116.688 registros (9,46%), seguido por Paraná (PR), com 84.341 (6,84%), e Santa Catarina (SC), com 49.702 (4,03%). A produção está relativamente bem distribuída entre os três estados da região. Já no Nordeste, com 209.049 registros (16,96%), destacam-se os estados de Pernambuco (PE), com 45.645 trabalhos (3,70%), Bahia (BA), com 41.438 (3,36%), e Ceará (CE), com 35.452 (2,88%), que concentram mais da metade da produção regional.

A Região Centro-Oeste, responsável por 86.692 trabalhos (7,03%), tem no Distrito Federal (DF) o maior volume, com 36.828 registros (2,99%), seguido por Goiás (GO), com 24.404 (1,98%). Mato Grosso (MT) e Mato Grosso do Sul (MS) apresentam participações menores, com 11.165 (0,91%) e 14.295 (1,16%) registros, respectivamente.

Por fim, a Região Norte, com o menor volume de produção (48.089 trabalhos, 3,90% do total), tem o Pará (PA) como principal destaque, com 24.991 registros (2,03%), seguido pelo Amazonas (AM), com 12.279 (1,00%). Os demais estados apresentam números bem mais baixos, refletindo a menor presença de programas de pós-graduação e infraestrutura científica consolidada. Essa distribuição evidencia uma possível concentração de esforços e investimentos em estados com maior densidade institucional, o que impacta diretamente o volume da produção acadêmica em nível nacional.

Passando agora para a distribuição dos trabalhos defendidos entre 2004 e 2022, de acordo com o grau acadêmico do discente, foram identificados quatro tipos de titulação: Mestrado, Mestrado Profissional, Doutorado e Doutorado Profissional, conforme mostra o Gráfico 16.

Gráfico 16 – Quantidade de trabalhos por grau de titulação, no período de 2004-2022

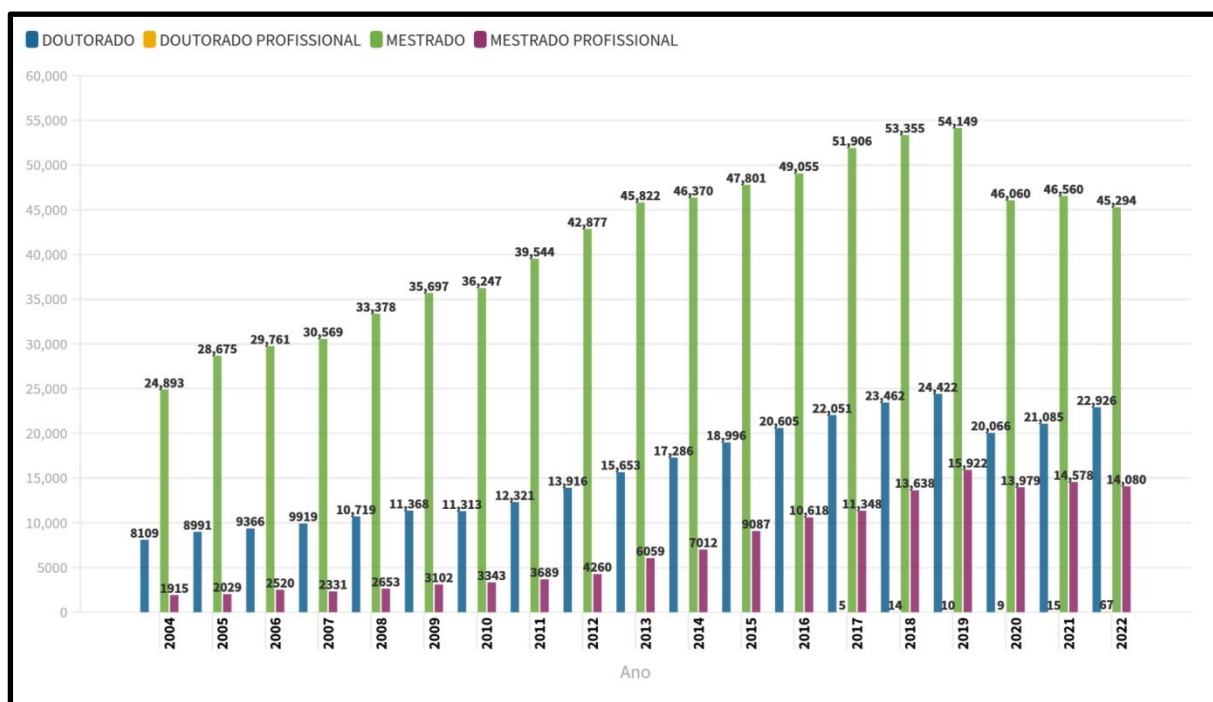


Fonte: Dados de pesquisa (2025).

Observa-se que o Mestrado concentra a maior parte da produção, com 788.013 trabalhos, o que representa 63,91% do total. Em seguida, o Doutorado contabiliza 302.574 registros, equivalente a 24,55%. O Mestrado Profissional totaliza 142.163 trabalhos, correspondendo a 11,53% da produção. O Doutorado Profissional aparece com apenas 120 registros, representando 0,01% do total no período analisado.

A evolução da produção acadêmica por grau de titulação, apresentada no Gráfico 17, confirma a predominância do Mestrado, responsável por 63,91% do total de trabalhos defendidos entre 2004 e 2022.

Gráfico 17 – Números de trabalho por grau em cada ano, no período 2004-2022

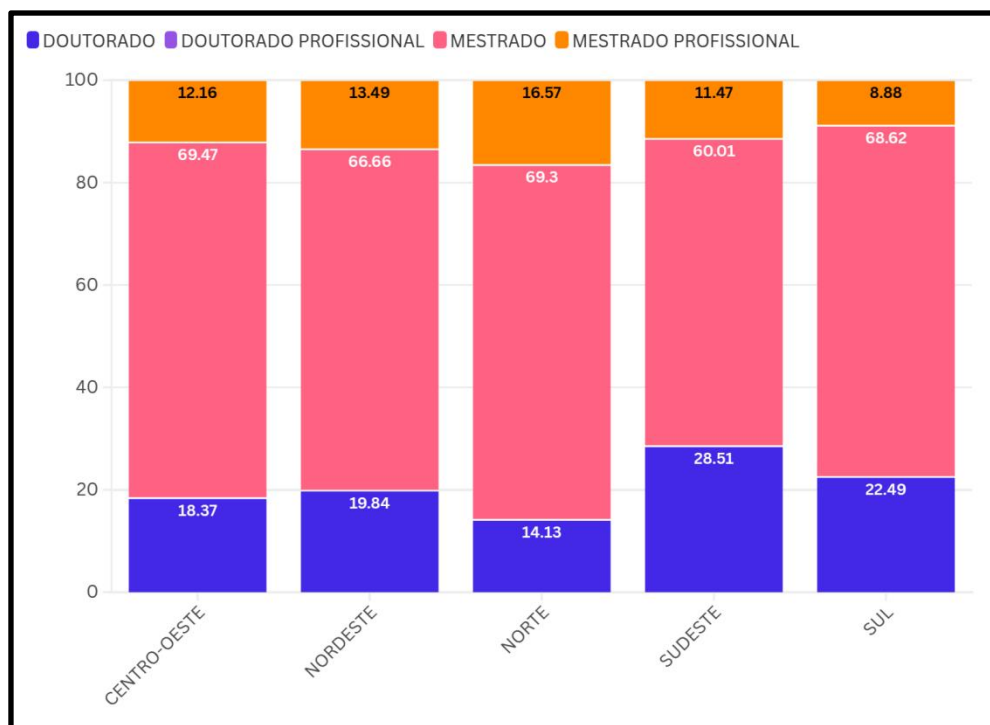


Fonte: Dados de pesquisa (2025).

O número de dissertações de Mestrado aumentou de 24.893 em 2004 para 54.149 em 2019, evidenciando a expansão da pós-graduação *stricto sensu* no Brasil. A partir de 2020, verificou-se uma queda para 44.207, provavelmente associada aos impactos da pandemia, com leve recuperação nos anos seguintes. O Doutorado, que representa 24,55% da produção total, também apresentou crescimento contínuo entre 2004 (8.109) e 2019 (24.422), com redução para 20.066 em 2020, seguida de recuperação moderada. O Mestrado Profissional, responsável por 11,53% dos trabalhos, registrou crescimento expressivo, passando de 1.915 dissertações em 2004 para 15.922 em 2019. Após a queda em 2020, os números estabilizaram-se em torno de 14.000 dissertações anuais, indicando consolidação da modalidade. O Doutorado Profissional, com participação residual de 0,01%, teve seus primeiros registros em 2017 e apresentou crescimento modesto, alcançando 67 trabalhos em 2022. A evolução dos dados confirma a consolidação do Mestrado e Doutorado acadêmicos como modalidades dominantes, ao passo que o Mestrado Profissional desponta como uma alternativa relevante voltada ao mercado.

A distribuição percentual dos graus acadêmicos por região, apresentada no Gráfico 18, confirma a predominância do Mestrado Acadêmico em todas as regiões do país, seguido pelo Doutorado e pelo Mestrado Profissional.

Gráfico 18 – Distribuição percentual dos graus acadêmicos por região, no período de 2004-2022



Fonte: Dados de pesquisa (2025).

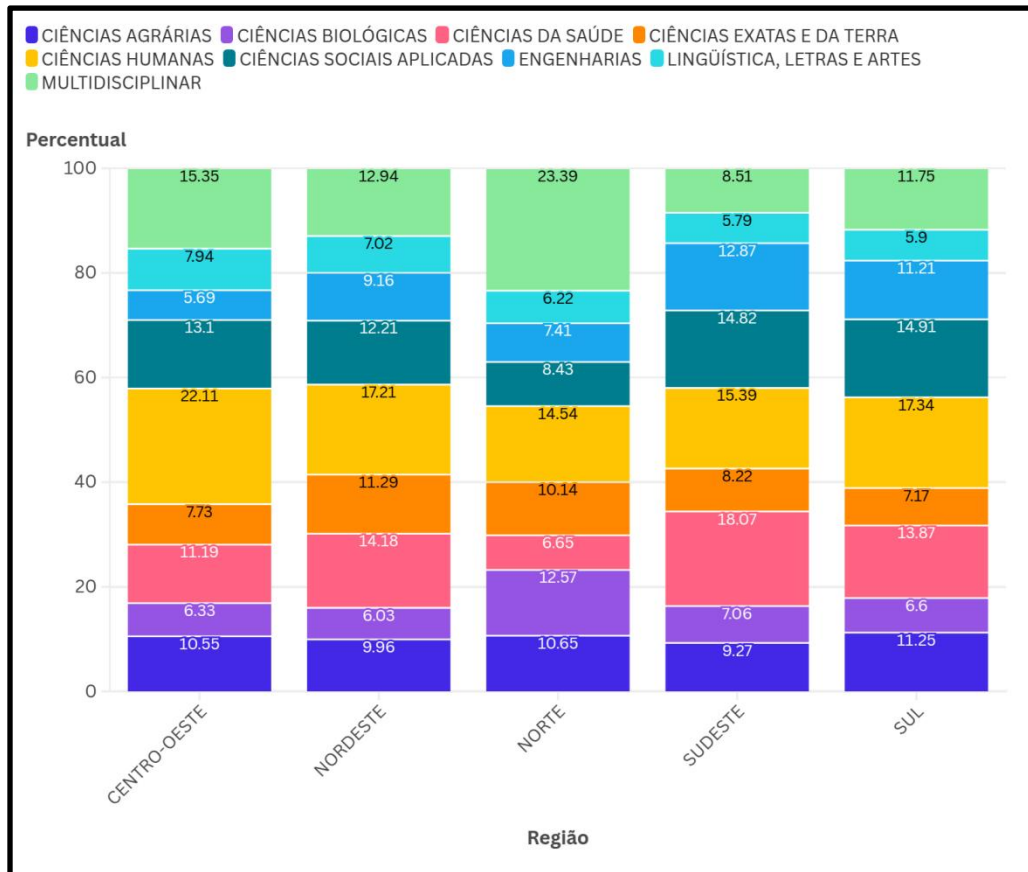
Analisando o Gráfico 18, observa-se que a Região Sudeste concentra os maiores volumes absolutos em todas as modalidades, com 383.056 dissertações de Mestrado (60,01%), 181.974 teses de Doutorado (28,51%) e 73.189 dissertações de Mestrado Profissional (11,47%), refletindo a estrutura consolidada da pós-graduação na região. A Região Sul apresenta 172.064 dissertações de Mestrado (68,62%), 56.398 teses de Doutorado (22,49%) e 22.256 dissertações de Mestrado Profissional (8,88%). O Nordeste registra 139.346 dissertações de Mestrado (66,66%), 41.480 teses de Doutorado (19,84%) e 28.208 dissertações de Mestrado Profissional (13,49%), sendo a segunda maior em volume dessa modalidade. No Centro-Oeste, há 60.222 dissertações de Mestrado (69,47%), 15.927 teses de Doutorado (18,37%) e 10.543 dissertações de Mestrado Profissional (12,16%). A Região Norte, com os menores volumes absolutos, apresenta 33.325 dissertações de Mestrado (69,3%), 6.795 teses de Doutorado (14,13%) e 7.967 dissertações de Mestrado Profissional (16,57%).

O Doutorado Profissional ainda é incipiente em todas as regiões, com destaque para o Sudeste, onde foram defendidos 90 trabalhos, seguido pelo Nordeste (15), Sul (13) e Norte (2). No Centro-Oeste, não há registros dessa titulação no período analisado. A presença pontual dessa modalidade indica que ela ainda não se consolidou no sistema de pós-graduação

brasileiro. Sua concentração no Sudeste pode estar associada à maior diversidade e quantidade de programas na região, conforme observado nos Gráficos 12 a 14.

A distribuição percentual das áreas do conhecimento por região, apresentada no Gráfico 19, evidencia grandes variações nos perfis regionais de produção acadêmica.

Gráfico 19 – Distribuição das áreas do conhecimento por região, no período 2004-2022



Fonte: Dados de pesquisa (2025).

Nota-se que a Região Sudeste apresenta os maiores percentuais em Ciências da Saúde, com 115.312 trabalhos (18,07%), Ciências Humanas, com 98.250 (15,39%), e Ciências Sociais Aplicadas, com 94.602 (14,82%). Há destaque também para Engenharias, com 82.130 registros (12,87%), e Ciências Exatas e da Terra, com 52.500 (8,22%). O padrão observado reflete a diversificação e consolidação da pós-graduação nessa região, que concentra instituições de grande porte e infraestrutura de pesquisa avançada.

A Região Sul apresenta perfil semelhante ao do Sudeste, com predomínio de Ciências Humanas, com 43.480 registros (17,34%), Ciências Sociais Aplicadas, com 37.381 (14,91%) e Ciências da Saúde, com 34.769 (13,87%). As Engenharias, com 28.108 (11,21%), também têm representação relevante, reforçando a presença de áreas tecnológicas na produção acadêmica regional.

A Região Nordeste concentra maior proporção de trabalhos em Ciências Humanas, com 35.984 (17,21%), Ciências da Saúde, com 29.644 (14,18%), e Ciências Sociais Aplicadas, com 25.526 (12,21%), mantendo também participação considerável em Ciências Exatas e da Terra, com 23.593 (11,29%), e nas Engenharias, com 19.155 (9,16%).

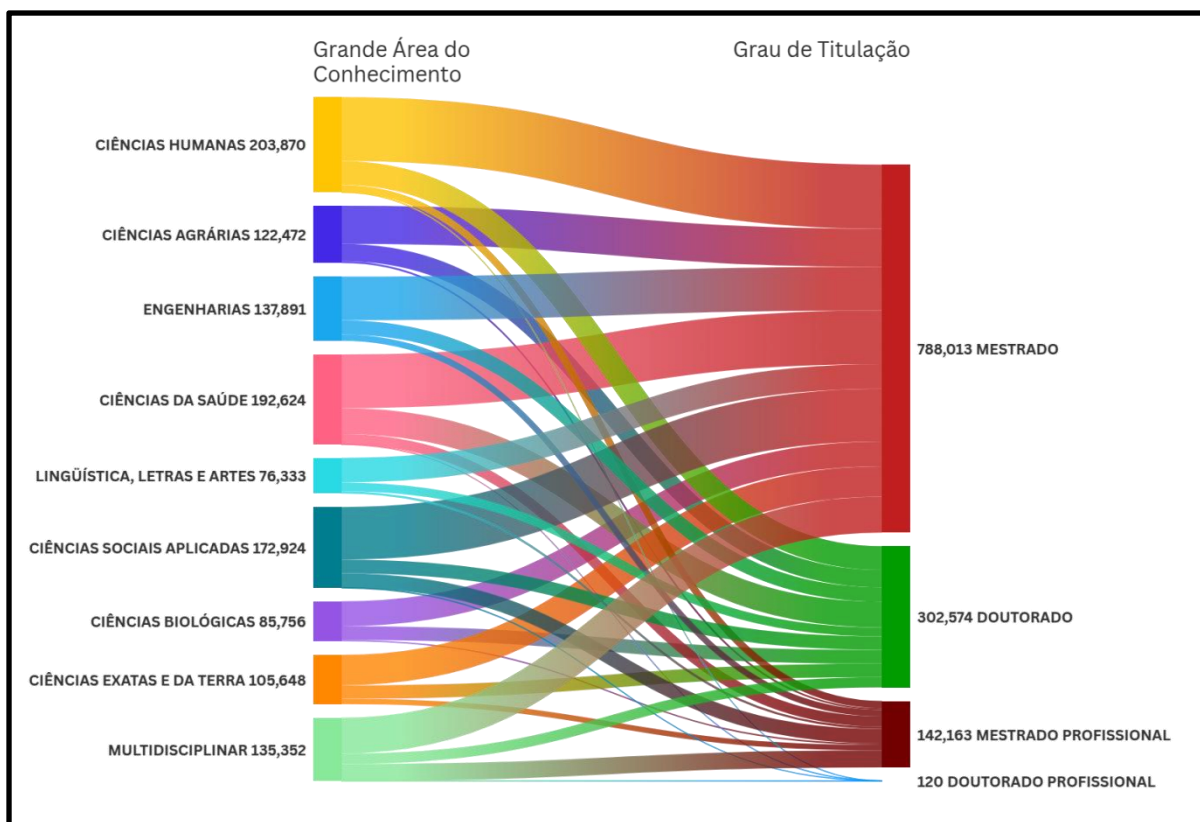
No Centro-Oeste, a distribuição segue um padrão semelhante ao do Nordeste, onde as Ciências Humanas predominam com 19.165 registros (22,11%). Em seguida, têm-se as Ciências Sociais Aplicadas, com 11.359 (13,1%). Apresenta menor volume em Ciências da Saúde, com 9.702 (11,19%), e Ciências Exatas e da Terra, com 6.700 (7,73%), em comparação com as regiões Sudeste e Sul. As Ciências Agrárias aparecem com 9.142 registros (10,55%), que é uma participação significativa nessa região.

A Região Norte apresenta o maior percentual de produção na área Multidisciplinar, com 11.245 registros (23,39%), o que sugere a predominância de programas que integram diferentes disciplinas para discutir assuntos em comum. Em seguida, se destacam as Ciências Humanas, com 6.991 (14,54%), muito próximas das Ciências Biológicas, com 6.042 (12,57%) das Ciências Agrárias, com 5.120 (10,65%), e das Ciências Exatas e da Terra, com 4.877 (10,14%). Com menores representatividades estão as Ciências da Saúde, com 3.197 (6,65%), e as Engenharias, com 2.991 registros (6,22%).

A distribuição das áreas do conhecimento por região está de acordo com os dados gerais da produção acadêmica nacional, conforme indicado no Gráfico 9, em que Ciências Humanas, Ciências da Saúde e Ciências Sociais Aplicadas concentram a maior parte dos trabalhos defendidos. A Região Sudeste, que lidera em volume total de produções, também apresenta os maiores percentuais em praticamente todas as áreas, refletindo sua boa infraestrutura de ensino e pesquisa. Por outro lado, a Região Norte exibe os menores valores em quase todas as áreas do conhecimento, o que evidencia a menor presença de instituições e programas de pós-graduação nessa localidade. Destaca-se a maior representatividade relativa das Ciências Agrárias nas Regiões Sul e Centro-Oeste, indicando a concentração de programas voltados a esse campo nessas regiões, possivelmente relacionada às vocações econômicas e produtivas locais.

Passando agora para os graus de titulação, a relação entre as grandes áreas do conhecimento com esses dados, apresentada no Gráfico 20, evidencia padrões distintos de distribuição entre Mestrado, Doutorado e Mestrado Profissional.

Gráfico 20 – Relação entre as grandes áreas do conhecimento e os graus de titulação, período 2004-2022



Fonte: Dados de pesquisa (2025).

Observa-se que as Ciências Humanas concentram o maior número de trabalhos, com 203.870 trabalhos, majoritariamente defesas de Mestrado (137.191), seguidas pelo Doutorado (51.135) e pelo Mestrado Profissional (15.539). Em seguida, as Ciências da Saúde somam 192.624 registros, sendo 115.174 dissertações de Mestrado, 55.259 teses de Doutorado e 22.185 dissertações de Mestrado Profissional. As Ciências Sociais Aplicadas, com 172.924 trabalhos, apresentam uma distribuição com destaque para o Mestrado (112.797), mas também com expressiva participação do Mestrado Profissional (31.079) e teses de Doutorado (29.025).

Nas Engenharias, com 137.891 registros, predominam as dissertações de Mestrado (93.685), seguidas pelas teses de Doutorado (30.765) e pelos Mestrados Profissionais (13.441). Já as Ciências Exatas e da Terra, com 105.648 trabalhos, têm perfil semelhante, com 65.156 dissertações de Mestrado, 28.605 teses e 11.887 dissertações profissionais.

As Ciências Agrárias, que totalizam 122.472 registros, têm predominância de Mestrados (81.478), seguidos pelos Doutorados (37.167), e com menor participação do Mestrado Profissional (3.827). A área Multidisciplinar, com 135.352 trabalhos, destaca-se pela maior presença relativa do Mestrado Profissional (36.258), além de 76.695 dissertações

de Mestrado e 22.313 teses. As Ciências Biológicas, com 85.756 registros, mantêm uma distribuição de maior número de Mestrados (53.349), seguidos pelos Doutorados (29.380), com menor presença do Mestrado Profissional (3.027). A área Linguística, Letras e Artes, com 76.333 registros, possui o menor volume total, apresentando 52.488 dissertações de Mestrado, 18.925 teses de Doutorado e 4.920 dissertações de Mestrado Profissional.

A comparação entre as áreas revela o predomínio do Mestrado em quase todas elas, especialmente nas Ciências Humanas, onde mais de dois terços dos trabalhos defendidos correspondem a dissertações. O Mestrado Profissional tem maior presença nas Ciências Sociais Aplicadas e na área Multidisciplinar, indicando a valorização de percursos formativos voltados à prática profissional e à aplicação do conhecimento em contextos não acadêmicos. Já nas Ciências Exatas e da Terra, Ciências Biológicas e Ciências Agrárias, observa-se uma maior paridade entre Mestrado e Doutorado, o que pode refletir um caráter mais consolidado da formação acadêmica nessas áreas, tradicionalmente vinculadas à pesquisa científica e tecnológica em laboratórios e de longo prazo. A menor representatividade do Doutorado é observada em Linguística, Letras e Artes, o que pode estar relacionado à menor oferta de programas doutorais sugerindo a necessidade de atenção em políticas de fomento à pesquisa avançada nessas áreas. O Doutorado Profissional, com apenas 120 registros em todo o período analisado, tem participação residual e, por isso, não é destacado na análise, embora apareça no gráfico. Sua adoção restrita a algumas áreas e instituições evidencia a necessidade de maior clareza quanto aos objetivos, reconhecimento e diretrizes dessa modalidade, que permanece à margem do sistema de pós-graduação *stricto sensu*.

O Gráfico 21, do tipo *sunburst*, apresenta a distribuição das titulações por grande área do conhecimento e por área de avaliação, segundo a classificação da CAPES (Arnold Reichert, 2023). Esse formato de gráfico permite visualizar relações hierárquicas, evidenciando como cada grande área se subdivide nas áreas de avaliação que a compõem, com base nas titulações registradas entre 2004 e 2022. Por limitações visuais, algumas áreas com nomes extensos, como “Administração Pública e de Empresas, Ciências Contábeis e Turismo” ou “Planejamento Urbano e Regional / Demografia”, não aparecem nomeadas no gráfico, mas estão representadas nas proporções internas do diagrama.

Gráfico 21 – Distribuição das 50 áreas de avaliação por grande área do conhecimento, período 2004-2022



Fonte: Dados de pesquisa (2025).

Na grande área das Ciências Exatas e da Terra, observa-se uma distribuição equilibrada entre as áreas, embora haja destaque para Química (27.362 titulações, 26,38%) e Ciência da Computação (26.258 titulações, 25,31%), que juntas concentram mais da metade da produção da área. Geociências (14.039 titulações, 13,53%), assim como Astronomia/Física e Matemática/Probabilidade e Estatística, os três com percentuais próximos a 17%, também apresentam participação expressiva. No restante, os percentuais variam entre 1,14% e 2,22% do total de titulações.

Na grande área das Ciências da Saúde, as maiores concentrações estão em Medicina I (45.544 titulações, 23,64%), seguida por Odontologia (32.834, 17,05%) e Saúde Coletiva (24.568, 12,75%), refletindo a centralidade dessas formações na pós-graduação da área.

A grande área Multidisciplinar, que representa 5,77% do total de titulações (71.108 registros), tem predominância da área Interdisciplinar, com 52,54% dos registros, seguida por Ensino (29.301, 21,65%) e Ciências Ambientais (17.908, 13,23%).

Na grande área das Ciências Sociais Aplicadas, três áreas concentram a maior parte das titulações: Administração Pública e de Empresas, Ciências Contábeis e Turismo, que concentram 52.187 titulações (30,18%), seguidas por Direito (49.713, 28,75%) e Comunicação e Informação (20.685, 11,96%), onde se encontra a Ciência da Informação.

Na grande área das Ciências Humanas, a Educação lidera com 78.653 titulações (38,58%). Em seguida vêm Psicologia (30.512, 14,97%), História (25.163, 12,34%) e Geografia (18.188, 8,92%).

Na grande área das Engenharias, observa-se uma distribuição relativamente equilibrada, com destaque para Engenharia III (43.050, 31,22%), seguida por Engenharia I (32.618, 23,65%) e Engenharia IV (31.819, 23,08%).

De forma geral, os resultados do conjunto DF3_trabalhos_BR revelam a expressiva expansão da produção acadêmica na pós-graduação *stricto sensu* entre 2004 e 2022, com destaque para o crescimento do Mestrado Profissional e a consolidação do Mestrado e do Doutorado acadêmicos. A análise evidencia desigualdades regionais, tanto em volume de trabalhos quanto na distribuição das áreas do conhecimento, com forte concentração nas regiões Sudeste e Sul e nas grandes áreas de Ciências Humanas, Ciências da Saúde e Ciências Sociais Aplicadas. O cruzamento entre áreas e titulações também permite identificar padrões formativos, como a maior incidência do Mestrado Profissional nas áreas Multidisciplinar e Sociais Aplicadas, bem como a paridade entre Mestrado e Doutorado nas Ciências Exatas, Biológicas e Agrárias. Por fim, a estruturação dos dados por área de avaliação, conforme a classificação da CAPES, oferece subsídios para compreender as prioridades formativas, as demandas institucionais e os prováveis focos de investimento na pós-graduação brasileira, compondo um panorama abrangente sobre a produção científica nacional no período analisado.

Na sequência, passa-se à análise do conjunto de dados DF4_programas_BR.

5.7.2 Resultados do conjunto de dados DF4_programas_BR

O conjunto de dados DF4_programas_BR, referente aos Programas de Pós-Graduação *stricto sensu* em funcionamento no Brasil, foi utilizado de forma complementar nesta pesquisa.

Seu principal papel foi servir como referência para a estruturação dos demais DataFrames, uma vez que contempla a maior parte dos campos presentes nos outros DataFrames utilizados.

Entretanto, sua cobertura temporal restringe-se ao período de 2013 a 2022, o que inviabiliza sua utilização para análises longitudinais abrangendo todo o intervalo estudado (2004–2022). Além disso, o elevado volume de dados associados aos programas de todas as áreas do conhecimento no país impôs limitações operacionais ao seu uso analítico detalhado no escopo deste trabalho, pois seria necessário analisar intelectualmente (manualmente) os dados, o que seria impraticável.

Diante disso, o DF4_programas_BR foi mobilizado pontualmente, sobretudo em consultas específicas ou em comparações localizadas, sempre com atenção às suas restrições de abrangência e granularidade. Ainda assim, sua inclusão contribui para complementar a análise institucional da pós-graduação brasileira, oferecendo subsídios para a compreensão da estrutura organizacional e das características gerais dos programas no período mais recente. Decidiu-se por não apresentar nesta seção a análise do conjunto brasileiro, devido a essas limitações. Porém, na seção 5.7.6 será feita uma análise mais aprofundada no conjunto DF4_programas_CI, que reúne os dados dos programas da área de Ciência da Informação. Tal recorte permitiu um tratamento mais detalhado das informações e maior coerência com os objetivos da pesquisa.

A seguir, a análise prossegue com o conjunto de dados DF5_docentes_BR, voltado à caracterização dos docentes da pós-graduação *stricto sensu* no Brasil.

5.7.3 Resultados do conjunto de dados DF5_docentes_BR

O conjunto de dados DF5_docentes_BR, referente aos docentes vinculados à pós-graduação *stricto sensu* no Brasil entre 2004 e 2022, reúne 1.482.959 registros, correspondentes a 123.591 docentes distintos. As variáveis analisadas contemplam informações como nacionalidade, ano de nascimento, categoria junto à instituição de ensino superior (IES), grau acadêmico, instituição de vínculo, ano e país da última titulação, além da área básica e da grande área do conhecimento associadas à formação do docente.

Esses campos possibilitam análises tanto individuais quanto cruzadas, permitindo identificar perfis de formação, trajetórias acadêmicas e padrões de distribuição dos docentes no sistema nacional de pós-graduação. A amplitude temporal e a diversidade de atributos do DF5 favorecem uma leitura mais detalhada sobre a qualificação e a inserção institucional do corpo docente, subsidiando investigações sobre a capacidade formativa das IES, a

especialização por áreas do conhecimento e a dinâmica de expansão da pós-graduação no período analisado.

Iniciando pela variável da nacionalidade, o Gráfico 22 apresenta a distribuição dos docentes da pós-graduação no Brasil no período de 2004 a 2022.

Gráfico 22 – Mapa com a nacionalidade dos docentes; período 2004-2022



Fonte: Dados de pesquisa (2025).

A grande maioria dos registros no Gráfico 22 é composta por docentes brasileiros, totalizando 117.724 indivíduos, o que representa 95,46% do total.

Por sua vez, os 5.565 registros de docentes estrangeiros correspondem a 4,54% do total de docentes distintos cadastrados no período. Entre os países estrangeiros com maior representatividade, destacam-se Portugal (514 registros, 0,42%), Argentina (508, 0,41%) e Estados Unidos (488, 0,40%), que juntos respondem por 25,98% dos docentes não brasileiros. Em seguida, aparecem França (457 registros, 0,37%), Espanha (399, 0,32%) e Itália (398, 0,32%), totalizando 22,06% desse grupo (docentes não brasileiros).

Outros países com destaque incluem Alemanha (354 registros, 0,29%) e Peru (282, 0,23%), que somados representam 11,43% dos docentes estrangeiros. Também se observam contribuições de Colômbia (229 registros, 0,19%), Cuba (159, 0,13%), Chile (155, 0,13%), Reino Unido (135, 0,11%) e Canadá (119, 0,10%), que, juntos, representam 13,81% do total de docentes internacionais.

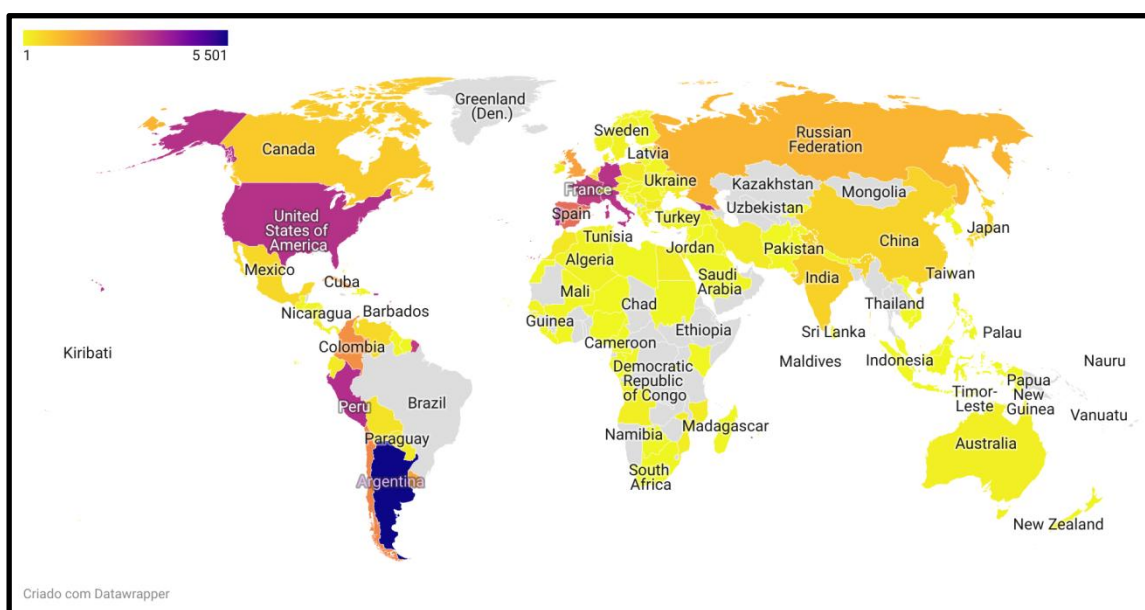
Em proporções menores, destacam-se países como Rússia (100 registros, 0,08%), México (85, 0,07%) e Índia (64, 0,05%), que totalizam 4,36% entre os docentes estrangeiros.

As demais nacionalidades apresentam participação inferior a 0,05% cada, compondo 13,36% dos registros de docentes não brasileiros.

A diversidade de nacionalidades, embora concentrada em alguns países e representando uma parcela reduzida do total, os docentes estrangeiros podem contribuir para a internacionalização da pós-graduação brasileira, indicando oportunidades de ampliação dos vínculos institucionais.

Complementando a análise da nacionalidade dos docentes da pós-graduação brasileira, o Gráfico 23 apresenta um mapa de calor com os registros de docentes estrangeiros entre 2004 e 2022, excluindo o Brasil da visualização para destacar os demais países.

Gráfico 23 – Mapa com a nacionalidade dos docentes sem o Brasil; período 2004-2022



Fonte: Dados de pesquisa (2025).

A visualização do Gráfico 23 evidencia que a América do Sul concentra a maior proporção de docentes estrangeiros na pós-graduação brasileira. Argentina lidera com 508 docentes (9,08% dos estrangeiros), seguida por Peru com 282 (5,04%) e Colômbia com 229 (4,09%). Também se destacam Cuba (159 – 2,84%), Chile (155 – 2,77%) e Uruguai (125 – 2,23%).

Na Europa, Portugal apresenta o maior número de docentes, com 514 registros (9,18%), seguido de França (457 – 8,17%), Espanha (399 – 7,13%), Itália (398 – 7,11%) e Alemanha (354 – 6,32%). Outros países europeus com participação relevante são Reino Unido (135 – 2,41%), Holanda (59 – 1,05%), Bélgica (51 – 0,91%), Áustria e Suécia, ambas com 17 registros (0,30%) cada.

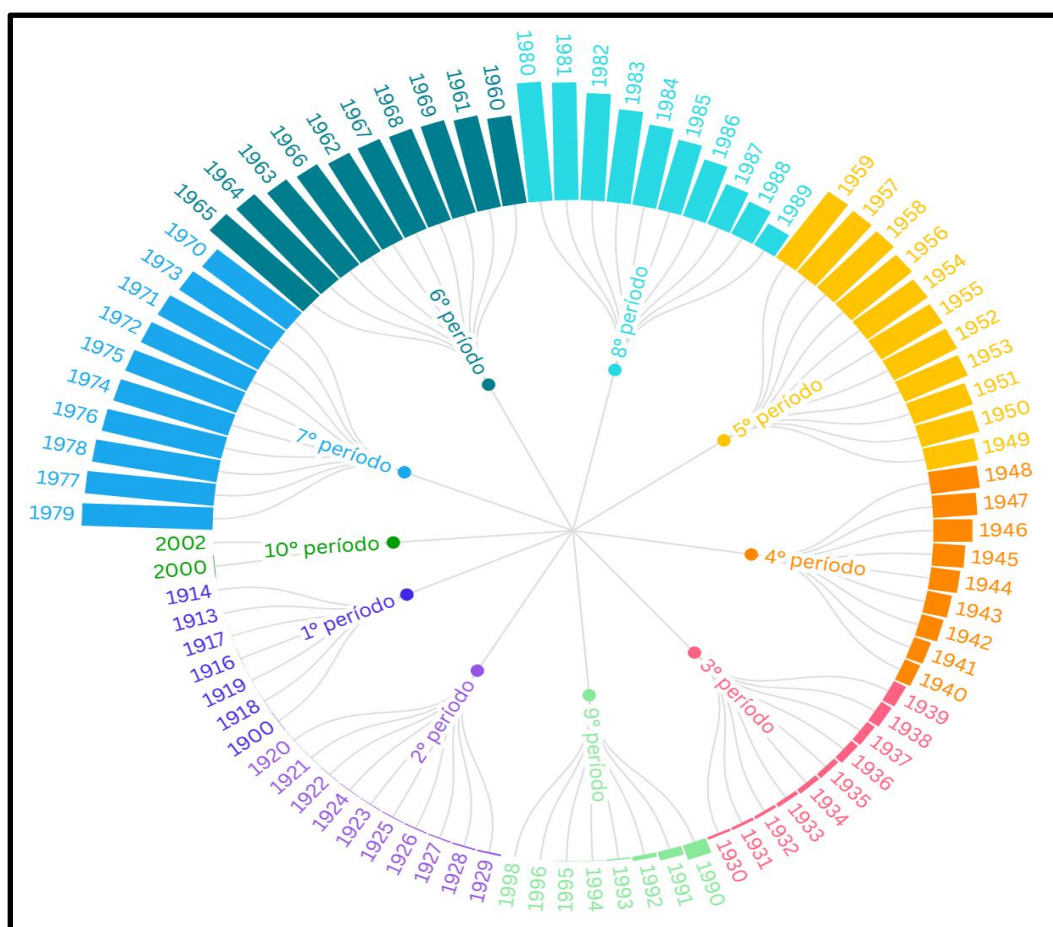
Na América do Norte, os Estados Unidos contabilizam 488 docentes (8,72%) e o Canadá, 119 (2,13%). Na Ásia, Índia destaca-se com 64 registros (1,14%), seguida por China (45 – 0,80%) e Japão (42 – 0,75%). No Oriente Médio, figuram o Irã (31 – 0,55%) e Israel (23 – 0,41%).

Quanto à África, os países com mais registros são Moçambique (23 – 0,41%), Angola (13 – 0,23%) e África do Sul (6 – 0,11%). Por fim, na Oceania, a Austrália registra 24 docentes (0,43%) e a Nova Zelândia, 9 (0,16%).

Esses dados reforçam a diversidade de vínculos internacionais da pós-graduação brasileira, com destaque para países de tradição acadêmica consolidada e laços históricos ou linguísticos com o Brasil. Ainda que a participação estrangeira seja modesta em termos proporcionais, ela revela uma base de cooperação científica que pode ser ampliada.

O gráfico 24, apresentado em formato radial e com agrupamento decenal (10 em 10 anos), exibe a distribuição dos anos de nascimento dos docentes da pós-graduação brasileira, permitindo identificar padrões de concentração etária ao longo do tempo.

Gráfico 24 – Distribuição dos anos de nascimento dos docentes; período 2004-2022



Fonte: Dados de pesquisa (2025).

Observa-se que o 7º período (1970–1979) concentra o maior número de docentes, totalizando 27.262 registros. Os anos com maior representatividade são 1979 (3.727 docentes – 3,02%), 1977 (3.681 – 2,98%) e 1978 (3.598 – 2,91%). Em seguida, o 6º período (1960–1969) reúne 32.596 docentes, destacando-se os anos de 1965 (3.663 – 2,96%), 1964 (3.573 – 2,89%) e 1963 (3.359 – 2,72%). Juntos, esses dois períodos compreendem a maioria dos docentes registrados.

No 8º período (1980–1989), nota-se uma tendência de declínio, com 23.468 docentes, sendo 1980 (3.705 – 3,00%), 1981 (3.653 – 2,96%) e 1982 (3.326 – 2,69%) os anos de maior frequência. O 9º período (1990–1999) registra uma queda acentuada, com apenas 952 docentes, dos quais 474 (0,38%) nasceram em 1990, com números decrescentes nos anos seguintes.

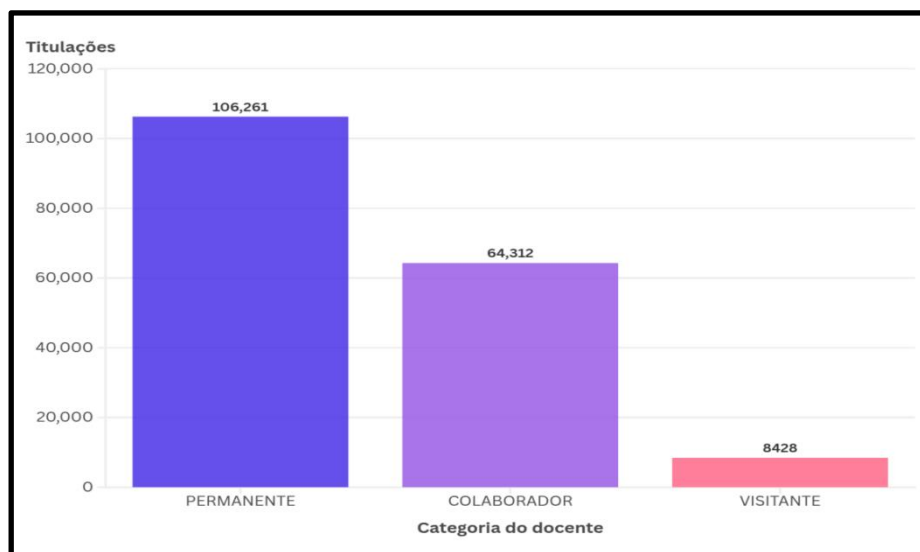
O 5º período (1950–1959) soma 21.451 docentes, com maior concentração em 1959 (2.527 – 2,04%), 1957 (2.472 – 2,00%) e 1956 (2.301 – 1,86%). Já o 4º período (1940–1949) contabiliza 9.953 docentes, com destaque para 1948 (1.400 – 1,13%), 1947 (1.262 – 1,02%) e 1946 (1.114 – 0,90%).

Nos períodos anteriores, observa-se redução progressiva nos registros. O 3º período (1930–1939) possui 2.381 docentes, com ênfase em 1939 (389 – 0,31%), 1938 (350 – 0,28%) e 1937 (272 – 0,22%). O 2º período (1920–1929) totaliza 245 registros, sendo 1929 (52 – 0,04%) o mais representativo. O 1º período (1900–1919) apresenta apenas 27 docentes, com destaque para 1900 (10 – 0,01%).

Há ainda registros pontuais que indicam possíveis inconsistências nos dados. Um docente aparece com idade de 0 anos e outro com 74.395 anos, ambos fora da plausibilidade cronológica. Além disso, há registros em 2000 (17 docentes – 0,01%) e 2002 (1 docente – 0,00%), o que pode sugerir casos de entrada precoce no sistema de pós-graduação ou erros no preenchimento dos dados. Não foi possível confirmar a veracidade desses registros por meio de fontes externas.

O Gráfico 25 apresenta a distribuição das categorias de docentes atuantes nos programas de pós-graduação *stricto sensu* no Brasil entre 2004 e 2022, com base nos registros da base DF5_docentes_BR.

Gráfico 25 – Distribuição das categorias dos docentes nas IES; período 2004-2022



Fonte: Dados de pesquisa (2025).

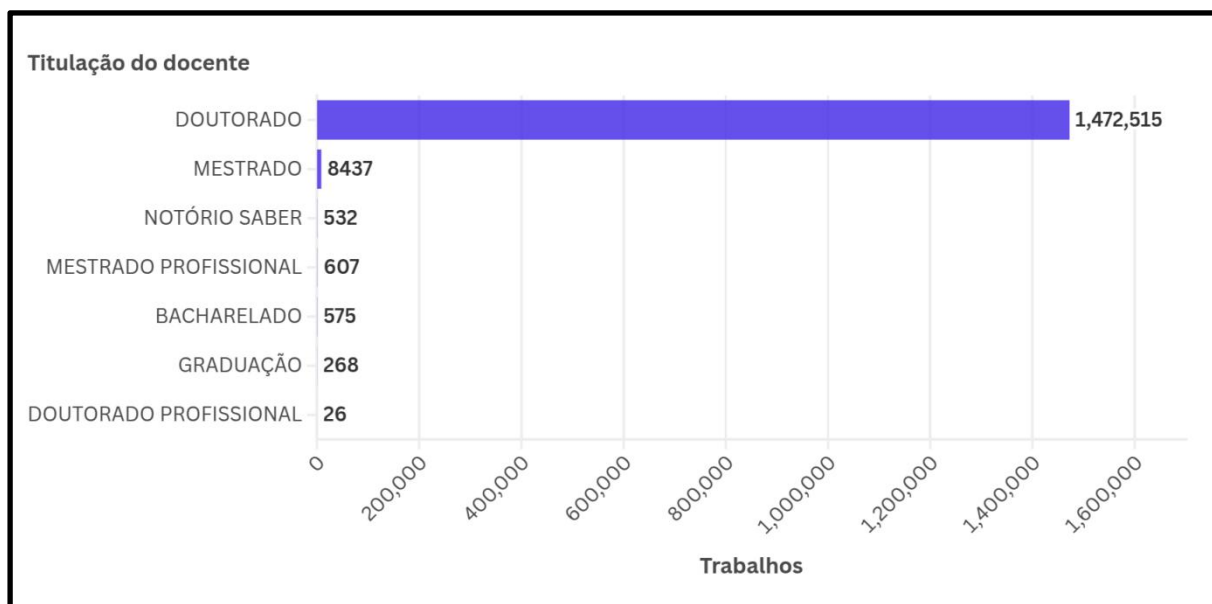
Observa-se que a maior representatividade está na categoria de docentes permanentes, com 106.261 registros, o que equivale a 59,36% do total. Esses docentes constituem a base principal dos programas, sendo responsáveis pela maior parte da orientação de dissertações e teses, atividades de pesquisa e pelas decisões acadêmicas nos colegiados, dando sustentabilidade dos programas de pós-graduação.

A categoria de colaboradores aparece em segundo lugar, com 64.312 registros (35,93%), indicando presença de docentes com participação parcial nas atividades dos programas. Esses docentes geralmente contribuem com disciplinas específicas, projetos de pesquisa e coorientações, mas sem vínculo integral com o programa. Importa destacar que, conforme as diretrizes da CAPES, os programas de pós-graduação devem manter um limite percentual de docentes colaboradores em relação ao número de docentes permanentes, justamente para assegurar a estabilidade e a coerência acadêmica do corpo docente.

Já a categoria de docentes visitantes, com 8.428 registros (4,71%), representa uma parcela menor da composição docente. Essa categoria inclui profissionais que atuam de forma temporária e geralmente são convidados por sua *expertise* para contribuir com disciplinas, bancas ou atividades pontuais, promovendo intercâmbio acadêmico entre instituições.

Passando agora para a distribuição do último grau de titulação dos docentes no momento da titulação dos discentes, apresenta-se o Gráfico 26.

Gráfico 26 – Distribuição dos títulos dos docentes; período 2004-2022



Fonte: Dados de pesquisa (2025).

A análise evidencia o forte predomínio do título de Doutorado, com 1.472.515 registros, o que corresponde a 99,30% do total. Esse número reforça o alinhamento dos programas de pós-graduação às exigências da CAPES, que estabelecem o doutorado como titulação padrão para docentes permanentes e orientadores.

Em seguida, aparecem os docentes com Mestrado, com 8.437 registros (0,57%), e aqueles titulados por Notório Saber, com 532 registros (0,04%) — uma modalidade tradicionalmente empregada em áreas como as artes, a cultura e os saberes tradicionais, onde trajetórias singulares e experiências práticas acumuladas têm grande valor formativo, que vem se estendendo para outras áreas do saber¹⁶ (Brasil, 2017a).

O Mestrado Profissional contabiliza 607 registros (0,04%), demonstrando sua presença ainda limitada na composição docente. Casos de docentes com Bacharelado (575 registros, 0,04%) e Graduação (268 registros, 0,02%) indicam condições pontuais, possivelmente ligados a supervisões técnicas ou orientações específicas autorizadas pelas instituições. Por fim, os docentes com Doutorado Profissional representam o menor grupo (26 registros, 0,00%), provavelmente por ser modalidade mais recente, reconhecido oficialmente pela CAPES em 2017 (Brasil, 2017b; Brasil, 2019) e ainda está em fase de consolidação.

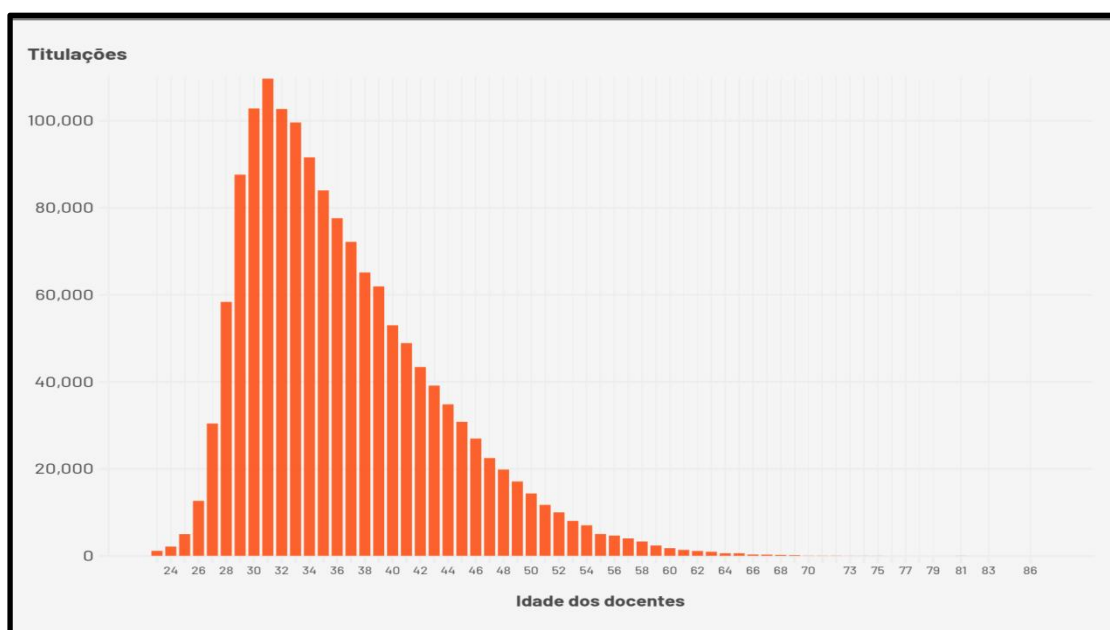
É importante destacar que esses dados representam o grau de titulação no momento da defesa do trabalho orientado — e não, necessariamente, a titulação atual do docente. Um

¹⁶ Notório saber instituído no inciso IV do Art. 6 da Lei nº 13.415/2017.

mesmo docente pode, por exemplo, orientar estudantes enquanto mestre e, mais tarde, obtido o título de doutor. Ainda assim, os dados sugerem uma forte aderência das instituições ao modelo acadêmico orientado à pesquisa, com exigência predominante de formação doutoral.

O próximo campo analisado refere-se ao ano da última titulação do docente. Embora esse dado, isoladamente, não permita identificar a idade do docente no momento da titulação, essa informação pôde ser inferida a partir da diferença entre o ano de nascimento do docente e o ano da defesa do trabalho do discente. Com base nesse cálculo, o Gráfico 27 apresenta a distribuição da idade dos docentes titulados no momento em que participaram da titulação dos discentes, permitindo observar os padrões etários predominantes na formação acadêmica ao longo do período analisado.

Gráfico 27 – Idade do docente no momento da última titulação; período 2004-2022



Fonte: Dados de pesquisa (2025).

A análise revela que a maior parte das titulações ocorre entre os 27 e 45 anos de idade, faixa que concentra mais de 70% do total de registros. A idade de 31 anos apresenta o maior pico de titulações, com 109.689 registros (7,41%), seguida de perto pelas idades de 30 (102.840; 6,95%) e 32 anos (102.725; 6,94%). As idades entre 33 e 35 anos mantêm volumes elevados, variando de 99.611 registros (6,73%) aos 33 anos, 91.589 registros (6,19%) aos 34 anos e 84.016 registros (5,68%) aos 35 anos. Essa concentração em torno dos 30-35 anos indica um padrão etário relativamente homogêneo e consolidado de titulação na pós-

graduação brasileira, alinhando-se ao percurso acadêmico tradicional da pós-graduação *stricto sensu*.

A partir dos 36 anos, observa-se uma tendência decrescente, mas ainda significativa, com 77.598 docentes titulados aos 36 anos (5,24%) e 61.941 aos 39 (4,19%). A curva segue em queda entre os 40 e 45 anos, período em que o número de titulações cai para menos da metade em relação ao pico, sendo 52.986 registros (3,58%) aos 40 anos, 48.930 (3,31%) aos 41 anos e 30.814 (2,08%) aos 45 anos.

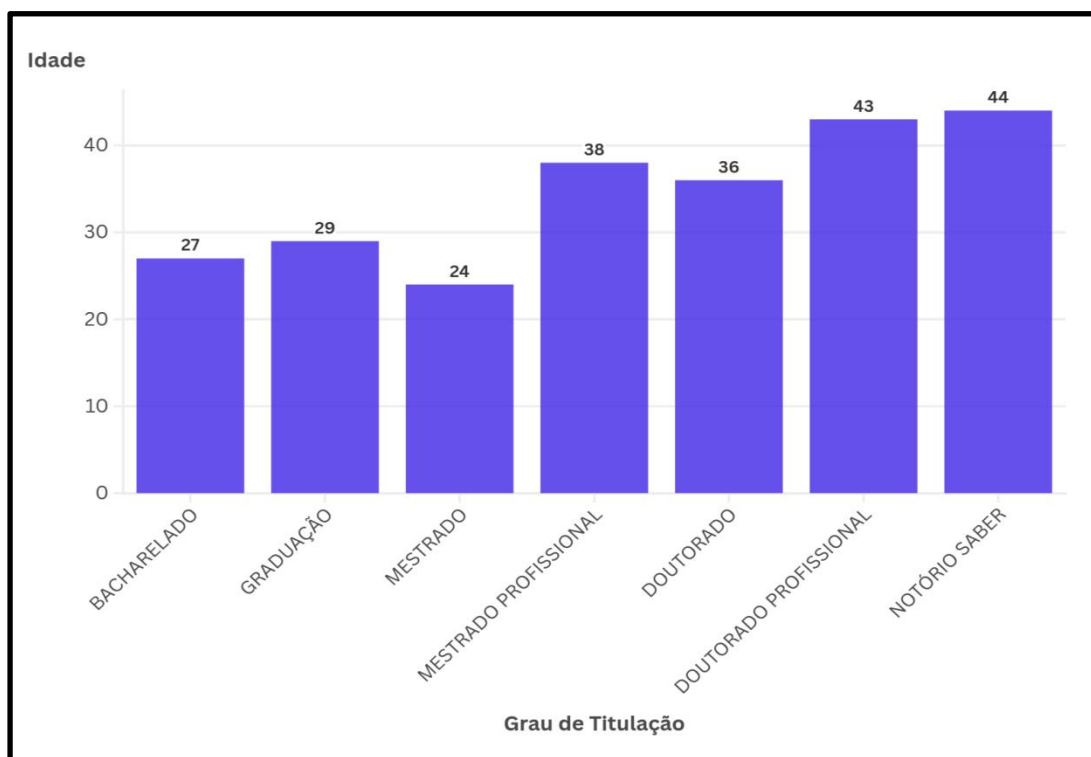
Entre 46 e 60 anos, a quantidade de docentes titulados continua a diminuir. Aos 46 anos, foram registrados 26.993 docentes (1,82%) A partir dos 55 anos, com 5.042 docentes (0,34%) titulados, os registros continuam caindo, com 4.023 (0,27%) aos 57 anos e 1.795 (0,12%) aos 60 anos.

Titulações em idades superiores a 60 anos tornam-se residuais: 625 registros (0,04%) aos 64 anos, 184 (0,01%) aos 69 anos e 68 (0,00%) aos 71 anos. Idades superiores a 75 anos possuem valores ainda menores, como 17 registros (0,00%) aos 76 anos, 10 (0,00%) aos 77 anos e apenas 2 registros (0,00%) aos 86 anos.

O gráfico não inclui registros com valores inconsistentes, como idades negativas, igual a 0, inferiores a 22 ou superiores a 89 anos. Dentre esses, destacam-se 9 registros com o valor 72.377 (possivelmente erro de digitação), 269 com idade igual a 0, e 2.649 com idade inferior a 22. Além disso, há 44 registros com idade superior a 89 anos. Essas ocorrências evidenciam a presença de ruídos nos dados, os quais, apesar de não invalidarem a análise geral, indicam a importância de mecanismos de validação nos sistemas de coleta e registro.

O Gráfico 28 apresenta a média de idade dos docentes no momento da obtenção de seu último grau de titulação.

Gráfico 28 – Distribuição da idade média por titulação dos docentes; período 2004-2022



Fonte: Dados de pesquisa (2025).

Os dados revelam que a menor média de idade dos docentes ocorre na titulação de Mestrado, com 24 anos, seguida por Bacharelado (27 anos) e Graduação (29 anos). No caso do Doutorado, a média é de 36 anos, enquanto no Mestrado Profissional esse valor sobe para 38 anos. A média de idade na obtenção do Doutorado Profissional é ainda mais elevada, situando-se em 43 anos, e a titulação por Notório Saber apresenta a maior média, com 44 anos — o que era esperado, considerando que se trata de um reconhecimento acadêmico baseado na trajetória da prática profissional.

A idade média de 24 anos para titulação em Mestrado, apresentada no Gráfico 28, contrasta com os dados do CGEE (2024), que indicam uma média superior ao longo do tempo. Segundo o Centro de Gestão e Estudos Estratégicos¹⁷ (CGEE), a idade média na titulação do Mestrado teve uma variação entre 33,4 anos (em 1996) e 33,8 anos (em 2021), com tendência de leve declínio entre 1996 e 2013 (32,2 anos), seguido por novo aumento. Esses valores, bem acima dos 24 anos apresentados na análise dos docentes, indicam que a média do Gráfico 28 pode refletir distorções oriundas da base de dados analisada, como erros de preenchimento, ou pela inclusão de docentes titulados em contextos excepcionais (por exemplo, profissionais com dupla titulação). Além disso, é possível que a média mostrada no Gráfico 28 tenha sido

¹⁷ O Centro de Gestão e Estudos Estratégicos é uma Organização Social vinculada ao Ministério da Ciência, Tecnologia e Inovação (MCTI), dedicada à realização de estudos estratégicos em ciência, tecnologia e inovação.

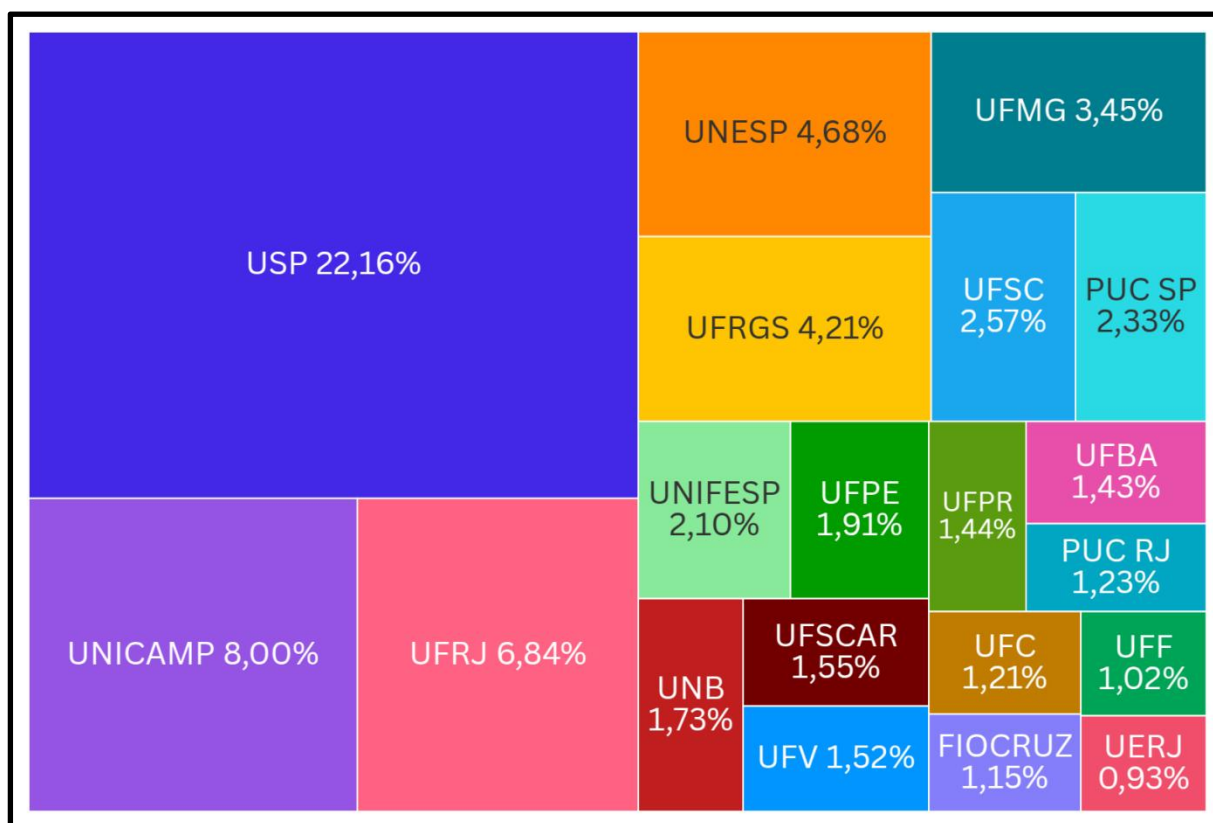
influenciada por um número restrito de registros com idades atipicamente baixas. Essa discrepância reforça a necessidade de cautela na leitura direta da média e ressalta a importância de considerar tanto a distribuição completa das idades quanto os dados de referência de estudos nacionais consolidados.

Para o doutorado, os dados do CGEE indicam uma trajetória de redução na idade média de titulação entre 1996 e 2019, passando de 39,4 para 37,0 anos. Essa tendência encontra um ponto de convergência com a média de 36 anos observada no Gráfico 28, sugerindo que muitos docentes obtiveram seus títulos durante o período de declínio identificado pelo CGEE. A leve elevação registrada nos anos de 2020 e 2021, quando a idade média voltou a subir para 37,4 anos, pode estar relacionada aos efeitos da pandemia da COVID-19 sobre o calendário acadêmico, o que temporariamente interrompeu a tendência de queda. A média de 36 anos observada no Gráfico 28, portanto, está bastante próxima do valor mais recente da série do CGEE, sugerindo consistência entre as bases de dados, especialmente no contexto da trajetória docente na pós-graduação. Esses dados reforçam a hipótese de que a média de idade entre docentes titulados reflete não apenas políticas estruturais de formação, mas também oscilações conjunturais, como emergências sanitárias ou mudanças institucionais.

Outros dois fatores importantes, que devem ser considerados, são, em primeiro lugar, que a base de dados aqui utilizada considera apenas docentes que atuaram na orientação de trabalhos de pós-graduação, o que tende a incluir aqueles com trajetórias mais aceleradas e formação mais precoce. Em segundo lugar, os dados do CGEE se referem à população geral de titulados no país, no período 1996-2021, enquanto os dados apresentados neste estudo referem-se apenas àqueles que atuaram como docentes no sistema de pós-graduação *stricto sensu*, no período 2004-2022.

Dando continuidade à análise da formação docente, o Gráfico 29 apresenta as vinte instituições de ensino superior (IES) com maior número de titulações, responsáveis por 71% do total de registros, o que permite identificar aquelas com maior influência na formação acadêmica de docentes da pós-graduação brasileira no período analisado.

Gráfico 29 – 20 IES com maior número de titulação entre os docentes; período 2004-2022



Fonte: Dados de pesquisa (2025).

Analisando o gráfico 29 a partir das cinco regiões brasileiras, tem-se:

- No Sudeste, a Universidade de São Paulo (USP) lidera com 328.552 titulações (22,16%), seguida pela Universidade Estadual de Campinas (UNICAMP), com 118.665 (8,00%), e pela Universidade Estadual Paulista (UNESP), que contabiliza 69.388 (4,68%). Ainda na região, a Universidade Federal do Rio de Janeiro (UFRJ-IBICT) acumula 101.386 titulações (6,84%), enquanto a Universidade Federal Fluminense (UFF) e a Universidade do Estado do Rio de Janeiro (UERJ) registram, respectivamente, 15.075 (1,02%) e 13.831 (0,93%). Em Minas Gerais, a Universidade Federal de Minas Gerais (UFMG) contabiliza 51.230 titulações (3,45%) e a Universidade Federal de Viçosa (UFV), 22.547 (1,52%).
- No Sul, a Universidade Federal do Rio Grande do Sul (UFRGS) apresenta 62.398 titulações (4,21%), enquanto a Universidade Federal de Santa Catarina (UFSC) e a Universidade Federal de São Carlos (UFSCar) somam, respectivamente, 38.097 (2,57%) e 23.042 (1,55%). Por sua vez, a Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS) registra 11.773 (0,79%).

- Na região Nordeste, a Universidade Federal de Pernambuco (UFPE) lidera com 28.255 titulações (1,91%), seguida pela Universidade Federal da Bahia (UFBA), com 21.249 (1,43%), e pela Universidade Federal do Ceará (UFC), com 17.968 (1,21%).
- No Centro-Oeste, a Universidade de Brasília (UnB) se destaca com 25.700 titulações (1,73%).
- Na região Norte, a Universidade Federal do Pará (UFPA) contabiliza 10.504 titulações (0,71%) e a Universidade Federal do Amazonas (UFAM), 1.829 (0,12%).

A análise regional da formação docente evidencia uma concentração no Sudeste, que se destaca como a principal região formadora de docentes da pós-graduação. O Sul apresenta também forte representatividade, com universidades federais consolidadas que contribuem fortemente para o total de registros. Já o Nordeste, embora com menor expressão numérica, demonstra avanços importantes, provavelmente refletindo políticas recentes de expansão da pós-graduação na região. O Centro-Oeste tem participação mais modesta, concentrada principalmente na capital federal, enquanto o Norte exibe a menor representatividade entre as regiões, sugerindo desigualdades estruturais na oferta e consolidação dos programas de pós-graduação *stricto sensu* no país.

No cenário internacional, ainda que o número de titulações obtidas no exterior represente uma fração modesta do total, observa-se a presença de instituições de reconhecido prestígio acadêmico, especialmente da Europa. Destacam-se cinco instituições com maior número de titulações de docentes: a University of London (8.036, 0,54%); a Université Paris Cité (4.293, 0,29%); a Université Pierre-et-Marie-Curie (hoje integrada à Sorbonne Université), (2.666, 0,18%); a University of Oxford (2.292, 0,15%); e a University of Manchester (2.285, 0,15%). Esse dado revela uma inclinação histórica de parte dos docentes brasileiros por centros de excelência localizados em países e instituições com forte tradição acadêmica e científica, que pode ser verificado em *rankings*¹⁸ internacionais utilizados para medir prestígio acadêmico, produção científica, impacto internacional e qualidade do ensino.

Dando continuidade à análise da formação dos docentes da pós-graduação brasileira, o Gráfico 30 amplia a perspectiva ao abordar o contexto internacional. Representado por um

¹⁸ Verificar em: QS World University Rankings: <https://www.topuniversities.com/university-rankings/world-university-rankings>; Times Higher Education (THE) World University Rankings: <https://www.timeshighereducation.com/world-university-rankings>; Academic Ranking of World Universities (ARWU): <http://www.shanghairanking.com/rankings/arwu>.

mapa de calor, ele exibe a distribuição geográfica dos países onde os docentes obtiveram sua última titulação entre 2004 e 2022, permitindo visualizar a inserção global dos vínculos formativos que compõem o sistema nacional.

Gráfico 30 – Mapa com países da última titulação do docente; período 2004-2022

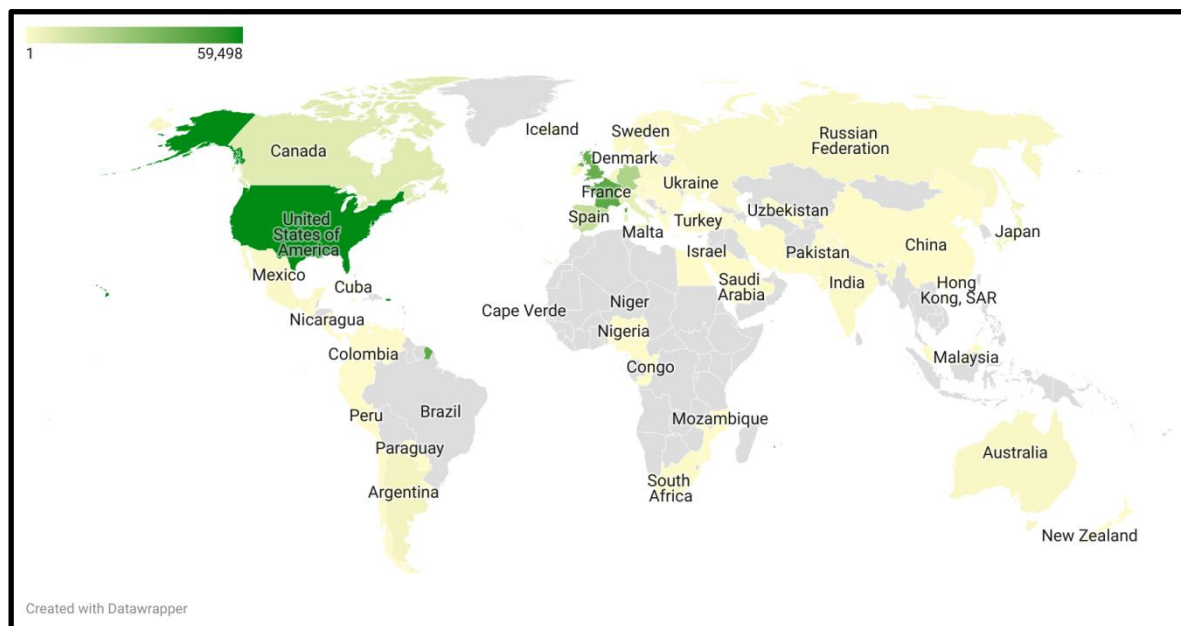


Fonte: Dados de pesquisa (2025).

A análise dos dados revela que 84,58% das titulações (1.254.273 registros) foram obtidas no Brasil, o que se reflete na coloração intensificada atribuída ao país no mapa. Em contraste, os demais países apresentam tonalidades mais suaves, evidenciando uma menor expressividade na formação de docentes no exterior. Apesar disso, a diversidade de países representados aponta para a presença, ainda que pontual, de trajetórias formativas internacionais.

Para aprofundar essa perspectiva, o Gráfico 31 destaca exclusivamente os países estrangeiros onde os docentes concluíram sua última titulação, permitindo observar com maior clareza a distribuição das titulações obtidas fora do Brasil.

Gráfico 31 – Mapa com países da última titulação do docente sem o Brasil; período 2004-2022

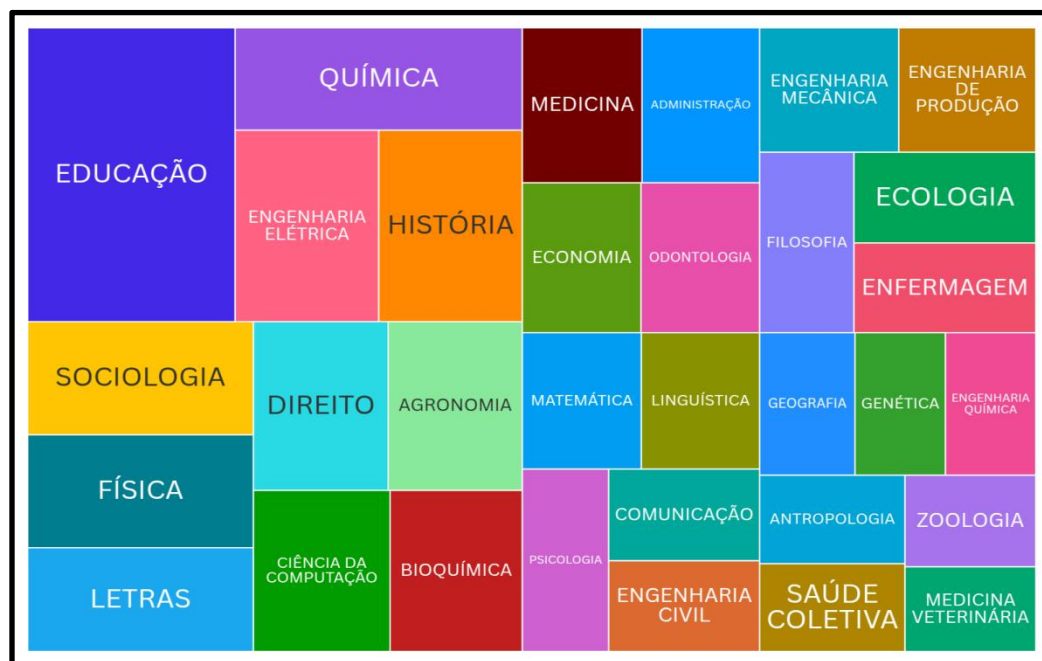


Fonte: Dados de pesquisa (2025).

Observa-se que os Estados Unidos lideram com 59.498 titulações (4,01%), seguidos por França (44.236 titulações, 2,98%), Reino Unido (41.276 titulações, 2,78%), Alemanha (21.167 titulações, 1,43%) e Espanha (15.871 titulações, 1,07%). Esses cinco países concentram quase 90% das titulações obtidas fora do Brasil, refletindo uma possível tendência de preferência. Canadá (8.343 titulações, 0,56%), Portugal (7.668 titulações, 0,52%), Itália (6.045 titulações, 0,41%), Bélgica (3.247 titulações, 0,22%) e Argentina (3.080 titulações, 0,21%) também aparecem entre os dez países mais representativos, ainda que com menor participação individual. As demais titulações no exterior — cerca de 22.761 registros (1,53%) — estão distribuídas entre países da Europa, América Latina, Ásia, África e Oceania, sugerindo uma diversidade geográfica, embora com frequência consideravelmente inferior.

Ampliando o escopo da análise, o Gráfico 32 apresenta a distribuição das áreas básicas de titulação dos docentes, permitindo observar as concentrações por campo do conhecimento no período analisado.

Gráfico 32 – Distribuição das áreas básicas da última titulação dos docentes; período 2004-2022



Fonte: Dados de pesquisa (2025).

Observa-se que as dez áreas com maior número de registros concentram uma parcela expressiva das titulações. A Educação se destaca como a área mais representativa, com 72.461 registros (4,89%), seguida por Química, com 35.082 registros (2,37%), Engenharia Elétrica, com 32.557 (2,20%), e História com 32.549 (2,19%), que possuem volumes bastante próximos. Áreas como Sociologia (30.295 registros, 2,04%), Física (30.209 registros, 2,04%), Letras (27.760 registros, 1,87%), Direito (26.942 registros, 1,82%), Agronomia (26.798 registros, 1,81%) e Ciência da Computação (26.063 registros, 1,76%), também figuram entre as mais recorrentes, todas com percentuais acima de 1,7%.

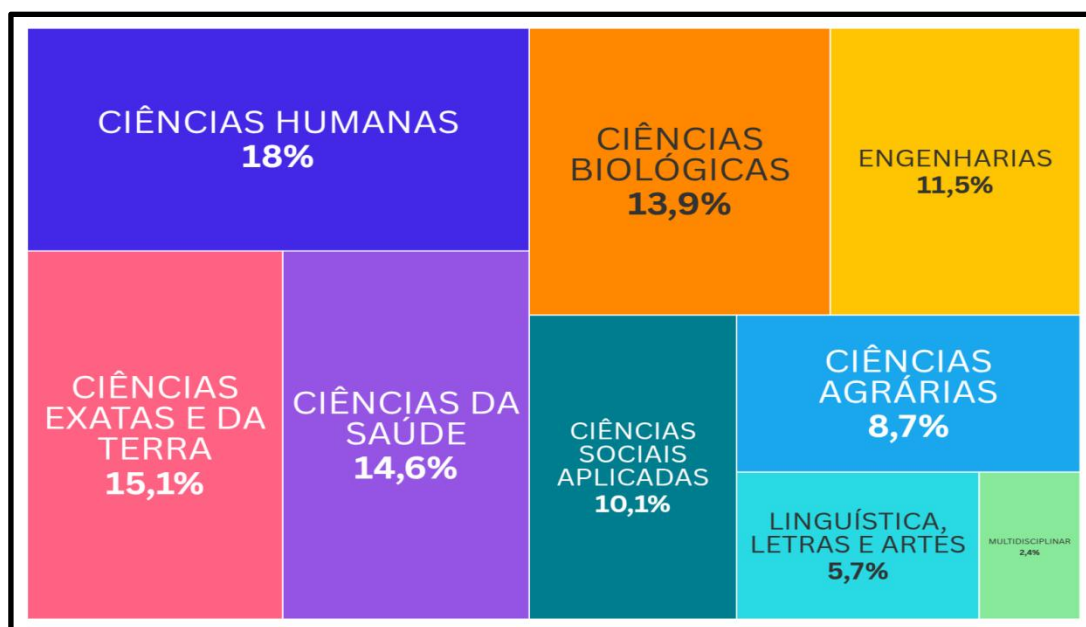
As demais áreas básicas de titulação, distribuídas entre 238 categorias distintas, revelam a amplitude do espectro disciplinar na formação docente. Entre essas, destacam-se Engenharia de Materiais e Metalúrgica (12.788 registros), Biologia Geral (12.696), Saúde Pública (12.378 registros), Fisiologia (11.853 registros), Botânica (11.478 registros), Ciência Política (11.472), Microbiologia (11.305 registros), Zootecnia (11.268 registros), Geociências (10.518 registros) e Farmacologia (10.226 registros), que, embora fora do grupo das mais frequentes, ainda representam volumes expressivos.

Esse cenário reflete a capilaridade e a especialização da pós-graduação brasileira, evidenciando a forte presença de áreas consideradas estratégicas pelo governo federal, como as Engenharias, Ciências Agrárias, Computação, Saúde, Educação e Meio Ambiente,

conforme estabelecido no PNPG 2024–2028(Brasil, 2023). Ao lado dessas, observa-se também a relevância de campos tradicionais das Humanidades e das Ciências Sociais, bem como a contribuição crescente de áreas emergentes e interdisciplinares, que ampliam a diversidade e a capacidade de resposta da pós-graduação aos desafios científicos, sociais e econômicos do país.

Dando continuidade à análise das titulações dos docentes, o Gráfico 33 consolida a distribuição por grandes áreas do conhecimento, permitindo uma visão agregada dos campos científicos predominantes na formação docente da pós-graduação brasileira.

Gráfico 33 – Distribuição das grandes áreas do conhecimento da última titulação dos docentes; período 2004-2022



Fonte: Dados de pesquisa (2025).

As Ciências Humanas lideram em número de registros, com 266.864 titulações (18,00%), evidenciando sua expressiva contribuição para a formação docente no país. Na sequência, destacam-se as Ciências Exatas e da Terra, com 223.926 registros (15,10%), e as Ciências da Saúde, com 215.811 (14,55%), ambas refletindo áreas de forte presença nos programas de pós-graduação. As Ciências Biológicas totalizam 205.976 titulações (13,89%), seguidas pelas Engenharias, com 170.830 registros (11,52%), indicando a relevância dessas áreas para o avanço científico e tecnológico.

As Ciências Sociais Aplicadas respondem por 149.857 registros (10,11%), enquanto as Ciências Agrárias reúnem 128.268 (8,65%), demonstrando sua importância para setores estratégicos como desenvolvimento regional e sustentabilidade. Já Linguística, Letras e Artes

contabilizam 84.816 registros (5,72%), e a área Multidisciplinar, 35.291 (2,38%). Além disso, 1.321 registros (0,09%) não possuem a grande área informada no banco de dados.

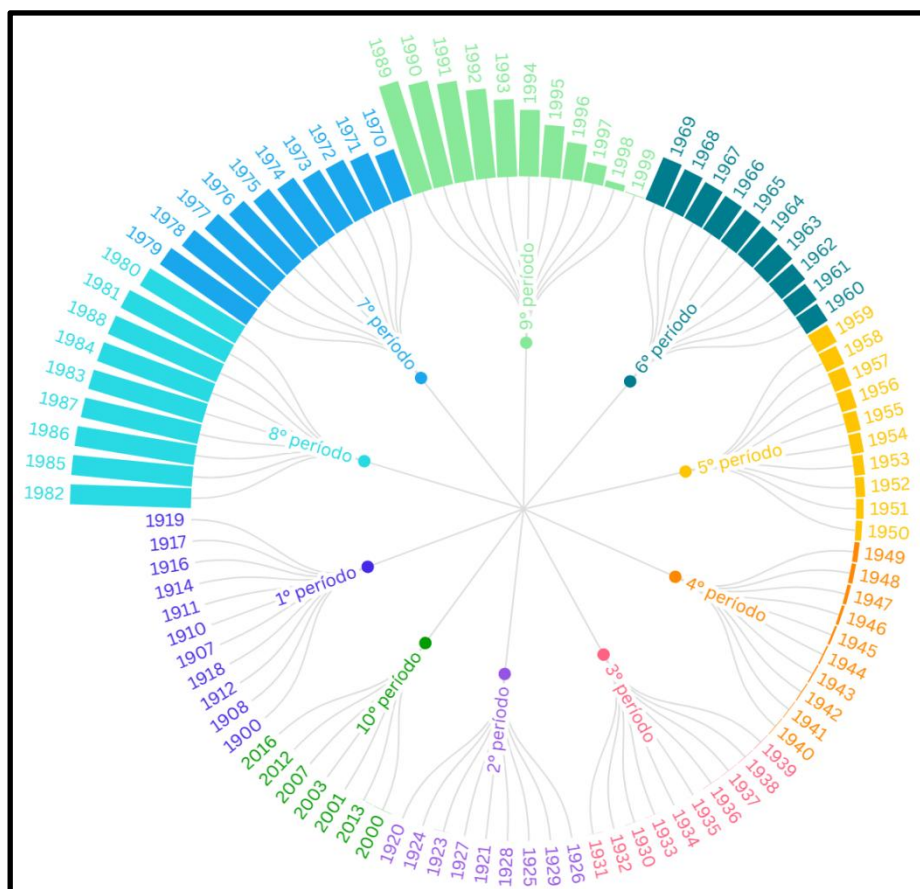
Encerrada a análise dos dados relativos aos docentes da pós-graduação *stricto sensu* no Brasil, a próxima seção (5.7.4) apresenta os resultados do conjunto de dados DF6_discentes_BR, voltando o foco para os discentes.

5.7.4 Resultados do conjunto de dados DF6_discentes_BR

Como parte da etapa de visualização dos dados, esta subseção apresenta a análise dos discentes da pós-graduação *stricto sensu* no Brasil, com base em 1.235.557 registros. As informações examinadas incluem ano de nascimento, país de origem, faixa etária, grau de titulação e tempo de conclusão do curso (em meses). Esses dados viabilizam análises tanto descritivas quanto combinadas, permitindo investigar os perfis formativos e as trajetórias acadêmicas dos discentes ao longo do período considerado.

O Gráfico 34 apresenta a distribuição dos anos de nascimento dos discentes da pós-graduação *stricto sensu* no Brasil entre 2004 e 2022. Para facilitar a análise, os dados foram organizados em dez períodos decenais: 1º período (1900–1919), 2º (1920–1929), 3º (1930–1939), 4º (1940–1949), 5º (1950–1959), 6º (1960–1969), 7º (1970–1979), 8º (1980–1989), 9º (1990–1999) e 10º período (2000–2016).

Gráfico 34 – Distribuição dos anos de nascimento dos discentes; período 2004-2022



Fonte: Dados de pesquisa (2025).

A segmentação estabelecida no Gráfico 35 permite observar a concentração de nascimentos em diferentes gerações que integraram a pós-graduação ao longo do período analisado.

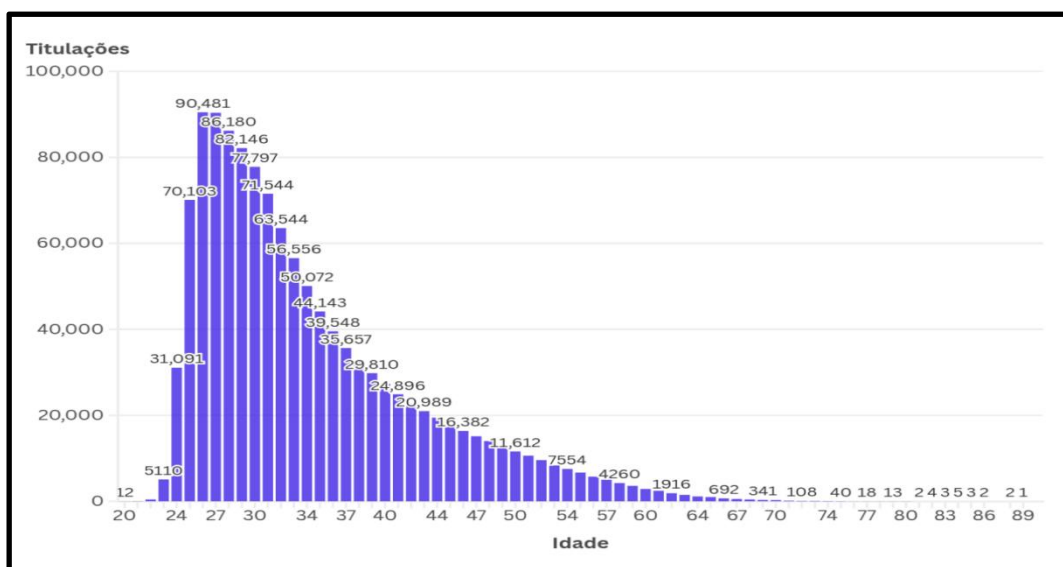
O 8º período (1980–1989) reúne a maior quantidade de registros, com 258.842 discentes. Dentro desse intervalo, os anos de 1985 (32.715), 1982 (32.722) e 1981 (31.912) são os mais representativos. Em seguida, o 7º período (1970–1979) contabiliza 205.431 registros, com destaque para 1979 (28.635), 1978 (26.767) e 1977 (24.919). O 9º período (1990–1999) totaliza 151.133 registros, com maior concentração em 1990 (28.855), 1991 (27.333) e 1992 (24.273). Já o 6º período (1960–1969) apresenta 99.543 registros, com picos em 1969 (12.447), 1968 (11.828) e 1967 (11.244).

Os períodos anteriores apresentam frequência decrescente. O 5º período (1950–1959) registra 36.752 discentes, enquanto o 4º período (1940–1949) reúne 6.317 registros. O 3º período (1930–1939) soma 436 registros, ao passo que os períodos 2 (1920–1929) e 1 (1900–1919) totalizam, respectivamente, 31 e 16 discentes, evidenciando o predomínio de gerações mais recentes na pós-graduação.

Destaca-se que foram identificados alguns dados inconsistentes, como 4 registros com ano de nascimento igual a 0 e registros entre 2000 e 2016 que, pela faixa etária, são improváveis e podem indicar erros de preenchimento ou inconsistência nos dados de origem.

Complementando a análise do perfil etário dos discentes, o Gráfico 35 aprofunda a investigação ao considerar a idade no momento da conclusão da pós-graduação, oferecendo uma perspectiva direta sobre o tempo de formação até a titulação.

Gráfico 35 – Distribuição da idade do discente no momento da titulação; período 2004-2022



Fonte: Dados de pesquisa (2025).

A abordagem utilizada nessa visualização permite estimar o tempo de formação até a conclusão do mestrado ou doutorado, oferecendo uma visão mais clara sobre o percurso dos discentes. Essas dez faixas etárias concentram a maioria das titulações registradas no período. Observa-se que a maior concentração de titulações ocorre entre 26 e 30 anos, faixa que representa o pico da distribuição. A idade mais frequente é 26 anos, com 90.481 registros (7,32%), seguida de 27 (7,31%) e 28 anos (6,97%). Essa tendência revela que a maior parte dos discentes conclui sua formação acadêmica *stricto sensu* ainda na faixa dos vinte anos. Até os 30 anos, as titulações mantêm altos volumes, como aos 29 (6,65%) e aos 30 anos (6,30%). A partir dos 31 anos, observa-se uma redução gradual nas titulações, mas ainda com representatividade: 31 anos (71.544, 5,79%), 32 (63.544, 5,14%), 33 (56.556, 4,58%), 34 (50.072, 4,05%) e 35 anos (44.143, 3,57%).

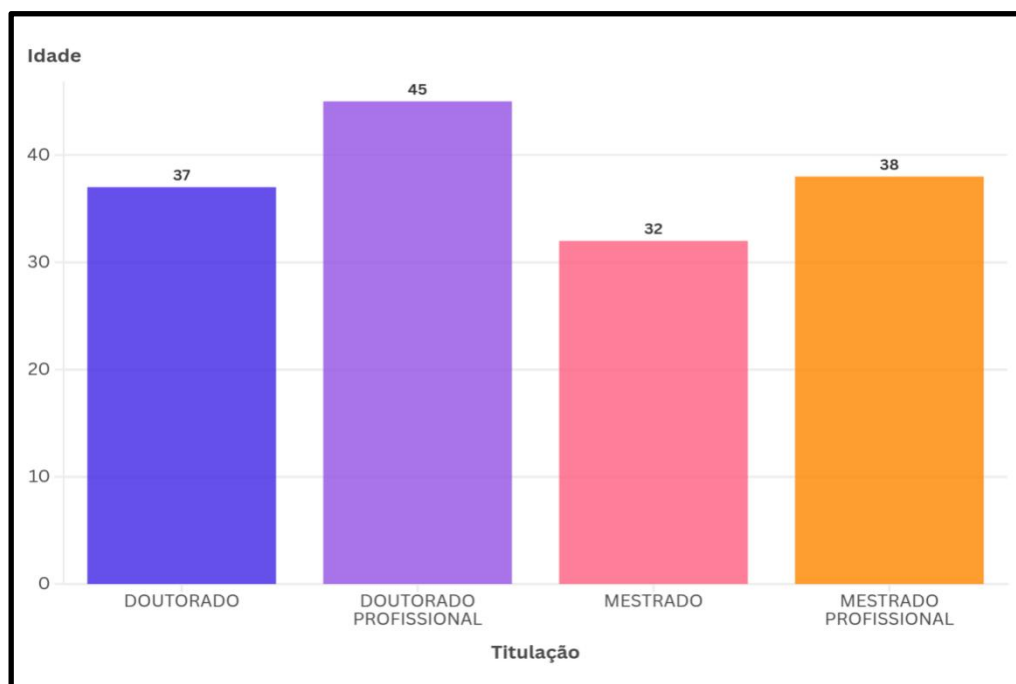
Após os 35 anos, a curva decresce de forma mais acentuada. Mesmo assim, há destaque para faixas entre 36 e 40 anos, que somam centenas de milhares de registros — como 36 anos (39.548, 3,20%), 37 (35.657, 2,89%), 38 anos (32.625 2,64%), 39 (29.810

2,41%) e 40 (27.288, 2,21%). A partir dos 40 anos, as titulações diminuem gradativamente, até os 50 anos (11.612, 0,94%), e seguem em queda para as idades superiores. Depois, as titulações tornam-se pontuais, embora ainda haja casos de conclusão aos 60 (1.916, 0,16%), 65 (692, 0,06%) e até 70 anos (341, 0,03%), refletindo trajetórias formativas atípicas ou prolongadas.

Ressalta-se que aqui também foram identificadas inconsistências nos dados: 49 discentes foram registrados com menos de 20 anos no momento da titulação, e outros 21 com idades superiores a 90 anos — valores que indicam prováveis erros de preenchimento ou falhas cadastrais. Esses casos foram mantidos apenas para fins de transparência na análise dos dados.

Após observar a distribuição etária no momento da titulação, o Gráfico 36 aprofunda a análise ao apresentar a idade média de conclusão da pós-graduação, segmentada por nível de titulação.

Gráfico 36 –Idade do docente no momento da última titulação; período 2004-2022



Fonte: Dados de pesquisa (2025).

Observa-se que a média de idade para conclusão do mestrado acadêmico é de 32 anos, enquanto o doutorado apresenta média de 37 anos, evidenciando um intervalo médio de cinco anos entre os dois níveis acadêmicos. No caso do mestrado profissional, a média é ligeiramente superior à do doutorado acadêmico, com 38 anos, com uma diferença de seis

com coloração mais suave, evidenciando uma inserção internacional ainda limitada nesse recorte.

Dando continuidade à análise da nacionalidade dos discentes, o Gráfico 38 apresenta a distribuição dos estudantes estrangeiros da pós-graduação no Brasil no período de 2004 a 2022, excluindo os registros de nacionalidade brasileira. A visualização, desenvolvida para complementar o Gráfico 37, permite identificar a representatividade de cada região do mundo por meio da análise por continente.

Gráfico 38 – Mapa com a nacionalidade dos discentes sem o Brasil; período 2004-2022



Fonte: Dados de pesquisa (2025).

A análise revela uma maior concentração de discentes estrangeiros oriundos da América do Sul, especialmente de países como Colômbia (4.538, 0,37%), Peru (2.843, 0,23%) e Argentina (882, 0,07%), evidenciando a relevância da cooperação regional na formação pós-graduada no Brasil. Depois vem Equador (496, 0,04%), Venezuela (437, 0,04%), Paraguai (384, 0,03%), Bolívia (345, 0,03%), Uruguai (269, 0,02%) e Chile (208, 0,02%).

Na América Central e Caribe, destacam-se Cuba (463, 0,04%), México (408, 0,03%), Honduras (141, 0,01%), Nicarágua (97, 0,01%) e Barbados (55, 0,00%). Já na América do Norte, destacam-se os Estados Unidos (222, 0,02%) e o Canadá (58, 0,00%).

Embora haja presença de discentes oriundos de todos os continentes, as contribuições da África lusófona e de países europeus e asiáticos permanecem numericamente mais discretas, indicando potenciais áreas para expansão de intercâmbios acadêmicos e políticas de internacionalização.

Os países africanos com maior número de discentes são Moçambique (800, 0,06%), Angola (346, 0,03%), Cabo Verde (209, 0,02%), Guiné-Bissau (164, 0,01%), Haiti (154, 0,01%), Congo (República Democrática) (86, 0,01%) e Namíbia (49, 0,00%).

Na Europa, destacam-se Portugal (378, 0,03%), França (299, 0,02%), Itália (321, 0,03%), Alemanha (328, 0,03%), Espanha (249, 0,02%), além de participações de países como Reino Unido (148), Ucrânia (102), Suécia (53), Dinamarca (48), Rússia (41), Malta (29) e Turquia (27).

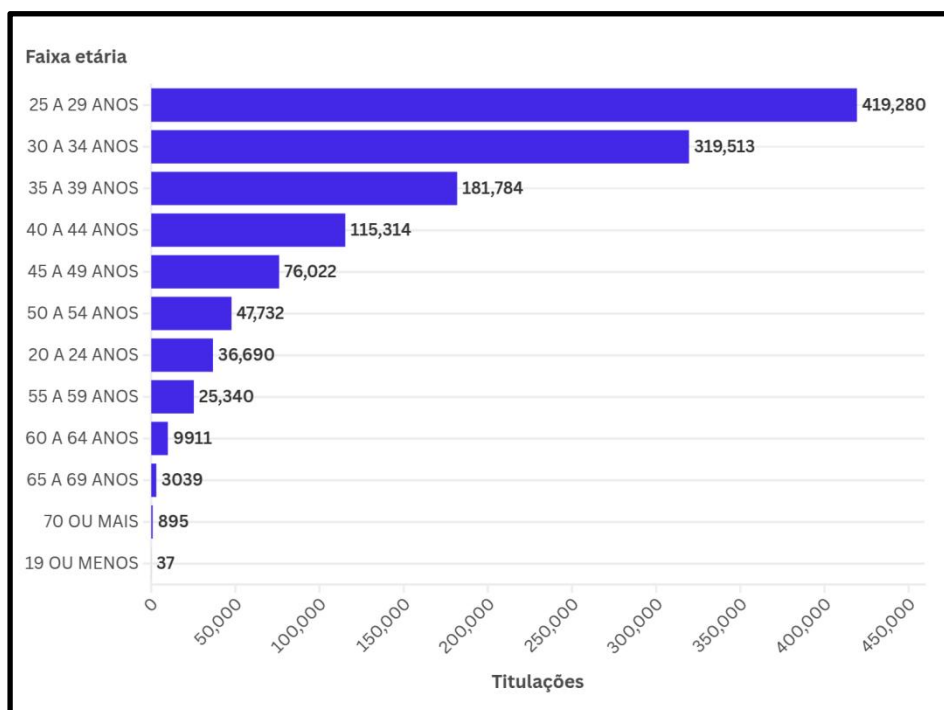
Na Ásia, destaca-se a China (124, 0,01%), seguida por Paquistão (105), Irã (97), Índia (94), Japão (58), Indonésia (44), Israel (31) e Arábia Saudita (29). Outros países com menor participação incluem Malásia, Sri Lanka, Mongólia, Cazaquistão e Uzbequistão.

Na Oceania, a Austrália aparece com 24 registros (0,00%), seguida pela Nova Zelândia, com 1 registro.

Embora representem apenas 1,42% do total de discentes, os estudantes estrangeiros evidenciam a diversidade de vínculos internacionais presentes na pós-graduação brasileira, com destaque para países da América Latina e da África lusófona. A predominância desses grupos está relacionada não apenas à proximidade geográfica e histórica, mas também à afinidade linguística, especialmente com os países de língua portuguesa, que encontram no Brasil um ambiente acadêmico mais acessível.

Dando prosseguimento à caracterização do perfil dos discentes, o Gráfico 39 apresenta a distribuição da faixa etária no momento da titulação, permitindo observar os padrões etários predominantes entre os concluintes da pós-graduação brasileira no período de 2004 a 2022.

Gráfico 39 – Faixa etária do discente da pós-graduação; período 2004-2022



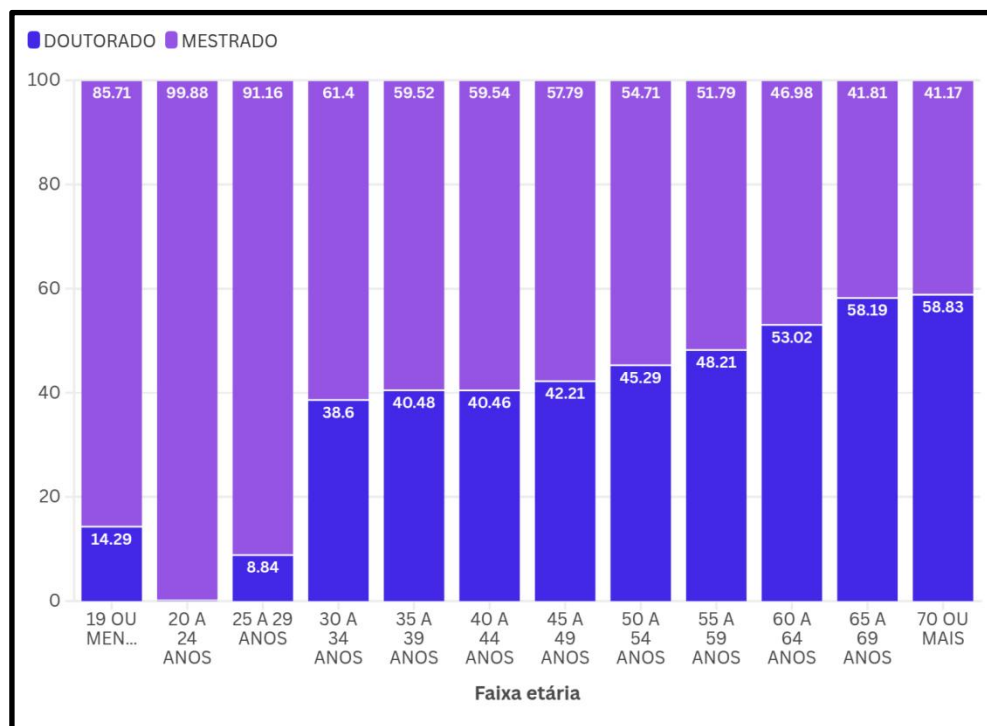
Fonte: Dados de pesquisa (2025).

A maior parte das titulações concentra-se entre 25 e 29 anos, com 419.280 registros (33,93%), seguida pelas faixas de 30 a 34 anos (319.513, 25,86%) e de 35 a 39 anos (181.784, 14,71%). Entre 40 e 44 anos, registram-se 115.314 titulações (9,33%), e entre 45 e 49 anos, 76.022 (6,15%). Nota-se que a participação diminui progressivamente a partir de 44 anos, ficando com menos de 10% dos registros em cada faixa superior. Esse padrão reflete uma tendência de ingresso e conclusão da pós-graduação mais cedo na carreira, com menor número de concluintes em fases mais avançadas da vida profissional.

Faixas etárias mais avançadas apresentam números progressivamente menores: 50 a 54 anos (47.732, 3,86%), 55 a 59 anos (25.340, 2,05%), 60 a 64 anos (9.911, 0,80%) e 65 a 69 anos (3.039, 0,25%). Entre os discentes com 70 anos ou mais, foram registradas 895 titulações (0,07%), enquanto aqueles com 19 anos ou menos somaram apenas 37 registros (0,00%).

Para aprofundar a análise etária, o Gráfico 40 detalha a distribuição das faixas etárias no momento da titulação, distinguindo entre os níveis de mestrado e doutorado, ambos acadêmicos.

Gráfico 40 – Comparação faixa etária discente doutorado e mestrado apenas acadêmico; período 2004-2022



Fonte: Dados de pesquisa (2025).

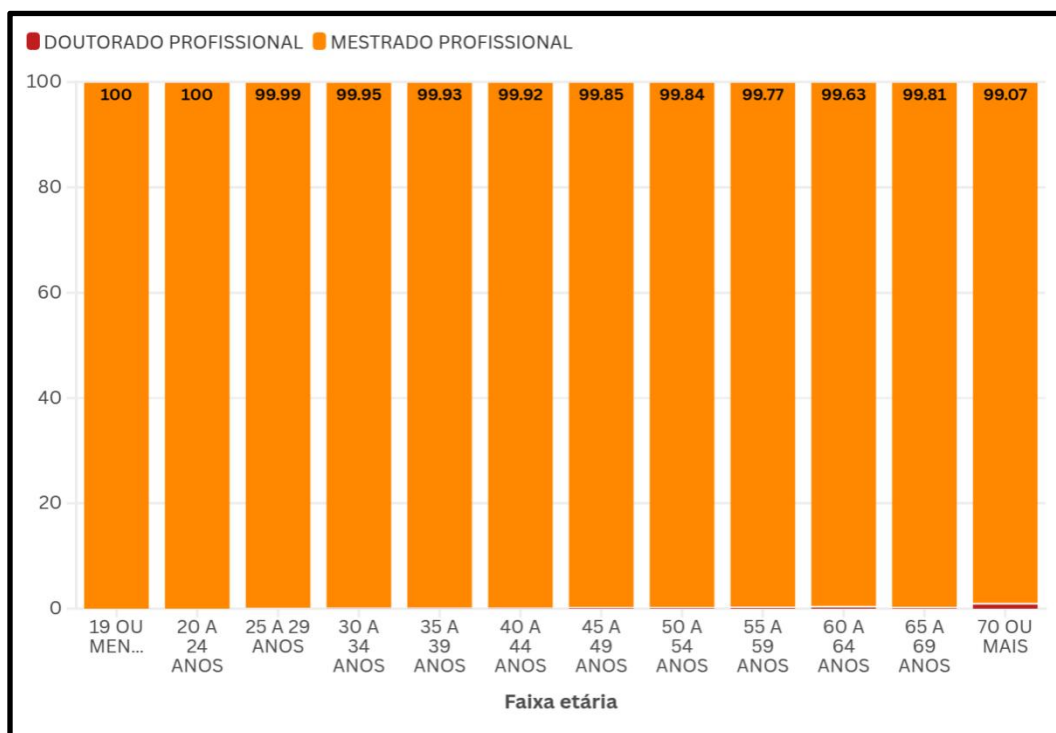
Os dados expostos no Gráfico 40 revelam que o mestrado acadêmico concentra a maior parte das titulações em faixas etárias mais jovens. A faixa de 25 a 29 anos registra 361.671 titulações (46,22%) no mestrado, enquanto o doutorado acadêmico contabiliza 35.059 (6,73%) nesse mesmo intervalo. Entre 30 e 34 anos, o mestrado acadêmico soma 175.270 registros (22,4%), e o doutorado acadêmico 110.204 (21,14%). A tendência se mantém nas faixas seguintes: de 35 a 39 anos, são 90.200 titulações (11,52%) no mestrado acadêmico e 61.347 (11,77%) no doutorado acadêmico; de 40 a 44 anos, os valores são 55.260 (7,06%) e 37.549 (7,2%), respectivamente.

A partir dos 45 anos, observa-se uma aproximação entre os dois níveis de titulação. Na faixa de 45 a 49 anos, o mestrado acadêmico registra 34.976 (4,47%) discentes e o doutorado acadêmico 25.551 (4,9%). Entre 50 e 54 anos, são 20.586 (2,63%) titulações no mestrado acadêmico e 17.044 (3,27%) no doutorado acadêmico; entre 55 e 59 anos, 10.383 (1,33%) e 9.667 (1,85%), respectivamente. Nas faixas de 60 a 64 anos, os valores se mantêm próximos (mestrado acadêmico: 3.767 - 0,48%; doutorado acadêmico: 4.252 - 0,82%), e essa tendência segue até 70 anos ou mais, com 463 (0,09%) registros no doutorado acadêmico e 324 (0,04%) no mestrado acadêmico. Um ponto a destacar é que o mestrado acadêmico também concentra

os poucos registros de discentes titulados com 19 anos ou menos (35 casos), contrastando com apenas 5 no doutorado acadêmico, reforçando o padrão de ingresso mais precoce nesse nível.

O Gráfico 41 dá continuidade a essa análise ao apresentar a distribuição etária dos discentes titulados no mestrado e doutorado, ambos profissionais.

Gráfico 41 – Comparação faixa etária discente doutorado e mestrado, apenas profissional; período 2004-2022



Fonte: Dados de pesquisa (2025).

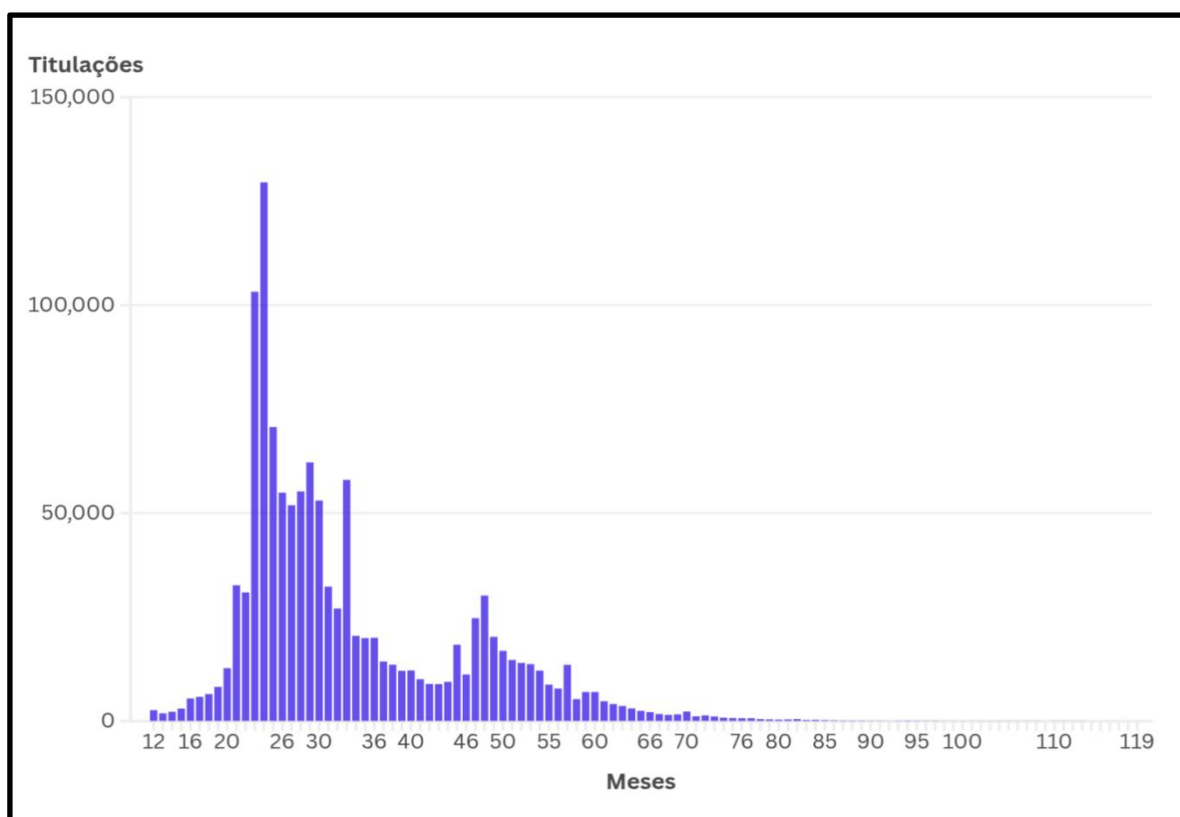
Observa-se que a titulação no mestrado profissional ocorre predominantemente em faixas etárias mais elevadas do que no mestrado acadêmico, confirmando o que vem sendo apresentado. A concentração se dá entre 30 e 39 anos, com 34.022 registros (23,13%) na faixa de 30 a 34 anos e 30.216 (20,55%) entre 35 e 39 anos. A partir dos 40 anos, os números diminuem de forma gradual, com 22.488 titulados (15,29%) na faixa de 40 a 44 anos e 15.472 (10,52%) entre 45 e 49 anos. Ainda assim, há presença de concluintes em faixas posteriores: 10.053 (6,84%) entre 50 e 54 anos, 6.369 (4,33%) entre 55 e 59 anos e 3.336 (2,27%) entre 60 e 64 anos. O mestrado profissional mantém titulações mesmo em idades avançadas, com 536 registros (0,36%) entre 65 e 69 anos e 107 (0,07%) com 70 anos ou mais.

Já o doutorado profissional apresenta um número bastante reduzido de registros, concentrando-se majoritariamente a partir dos 40 anos, refletindo o perfil mais maduro dos discentes. Entre 40 e 44 anos, foram identificados 17 registros (20,99%); entre 45 e 49 anos, 23 (28,40%), sendo esta a faixa mais representativa. As faixas seguintes mantêm números

decrecentes: 16 registros (19,75%) entre 50 e 54 anos, 12 (14,81%) entre 55 e 59 anos, 7 (8,64%) entre 60 e 64 anos, e 1 registro (1,23%) tanto entre 65 e 69 anos quanto na faixa de 70 anos ou mais. O número reduzido de titulados no doutorado profissional está diretamente relacionado à sua recente regulamentação. A modalidade foi instituída por meio da Portaria nº 389/2017 (Brasil, 2017), o que limita o volume de concluintes no período analisado.

O Gráfico 42 amplia essa investigação ao apresentar a distribuição dos meses utilizados para a titulação, sem, neste momento, separar os números entre mestrados e doutorados. Deve-se levar em consideração os parâmetros estabelecidos pela Resolução nº 5, de 10 de março de 1983¹⁹, que fixa a duração mínima de um ano para o mestrado e de dois anos para o doutorado.

Gráfico 42 – Distribuição da quantidade de meses para titulação; período 2004-2022



Fonte: Dados de pesquisa (2025).

Observa-se que a maior concentração de titulações se situa entre 23 e 24 meses, com 103.209 registros (8,35%) e 129.516 registros (10,48%), respectivamente, o que revela a predominância desse intervalo como tempo típico de conclusão de mestrado. Outros prazos próximos também se destacam, como 25 meses (5,72%), 26 meses (4,45%) e 27 meses

¹⁹ Disponível em: <https://cad.capes.gov.br/ato-administrativo-detallar?idAtoAdmElastic=233>. Acesso em: 8 mar. 2025

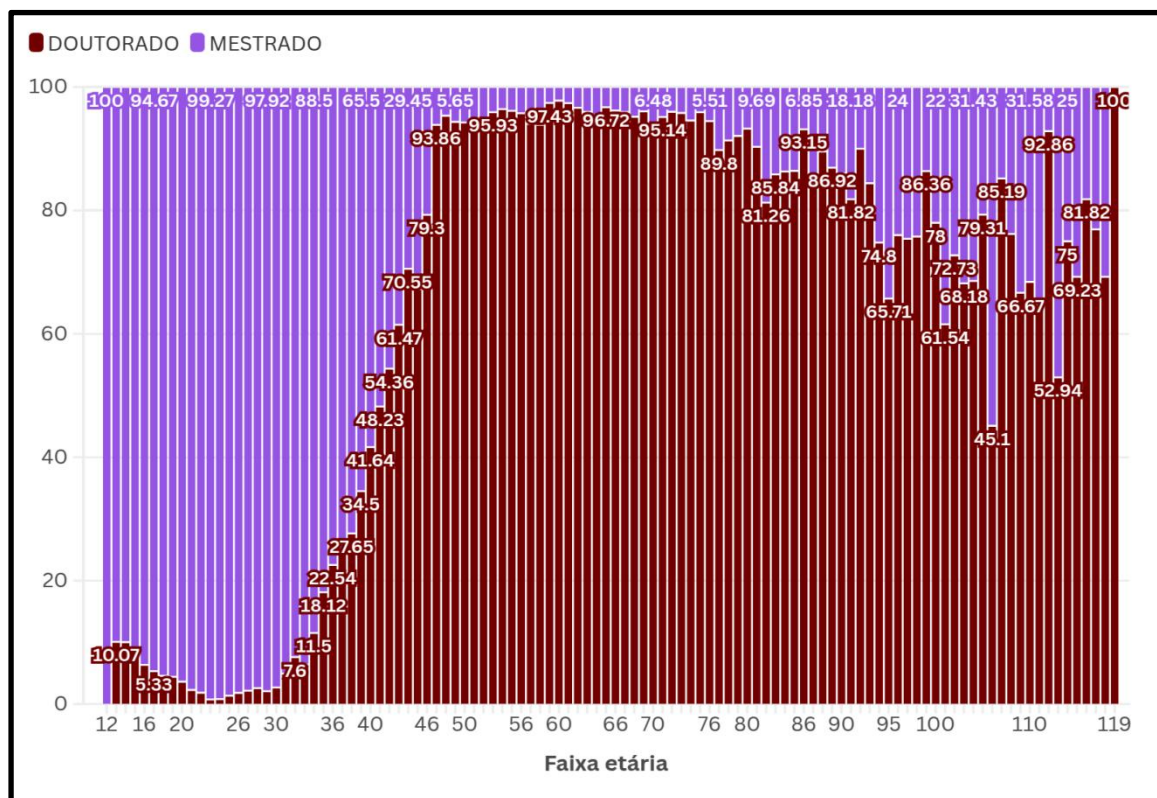
(4,20%), indicando variações que podem refletir diferenças entre programas, calendários institucionais ou fatores individuais. Há ainda distribuição expressiva entre 28 e 30 meses, com percentuais entre 4,29% e 5,03%, o que reforça a diversidade de prazos de conclusão observada no sistema.

Nos prazos superiores a 36 meses, observa-se uma mudança de padrão, com perfil de maior associação ao doutorado. No intervalo de 45 a 54 meses, destacam-se 45 meses (18.339 registros, 1,48%), 47 meses (24.717 registros, 2,00%) e, principalmente, 48 meses (30.176 registros, 2,44%), sugerindo uma concentração em torno de quatro anos — tempo comum de conclusão do doutorado. Entre 49 e 54 meses, os percentuais variam entre 1,64% e 0,98%, reforçando essa tendência. Após 60 meses, a frequência de titulações diminui gradualmente, embora ainda existam registros com mais de 100 meses, sinalizando casos excepcionais de prorrogação ou interrupção parcial do percurso acadêmico.

É importante observar que 12.890 registros (1,04%) indicam titulações em menos de 12 meses, e 112 discentes (0,01%) ultrapassaram 120 meses. Esses dados não foram representados no gráfico por destoarem dos limites normativos e acredita-se estarem relacionados a inconsistências cadastrais ou a situações excepcionais no histórico acadêmico dos discentes.

Para aprofundar a análise do tempo de conclusão, o Gráfico 43 complementa o anterior ao desagregar os dados por nível de titulação, permitindo comparar a distribuição dos meses para titulação entre o mestrado e o doutorado, ambos acadêmicos.

Gráfico 43 – Comparação meses para titulação dos discentes doutorado e mestrado acadêmico; período 2004-2022



Fonte: Dados de pesquisa (2025).

Os dados revelam que o mestrado acadêmico concentra a quase totalidade das titulações nos períodos iniciais. O tempo de 24 meses apresenta o maior volume, com 112.747 (34,6%) registros, seguido por 23 meses (89.921, 27,6%) e 25 meses (59.556, 18,3%), configurando o padrão esperado conforme a duração prevista na regulamentação. A partir de 30 meses, observa-se uma queda acentuada nos registros de mestrado acadêmico, indicando que a maioria das titulações ocorre dentro do prazo estipulado de dois anos. Assim, no total, cerca de 80% das titulações de mestrado ocorrem entre 23 e 30 meses, indicando forte adesão ao cronograma estipulado.

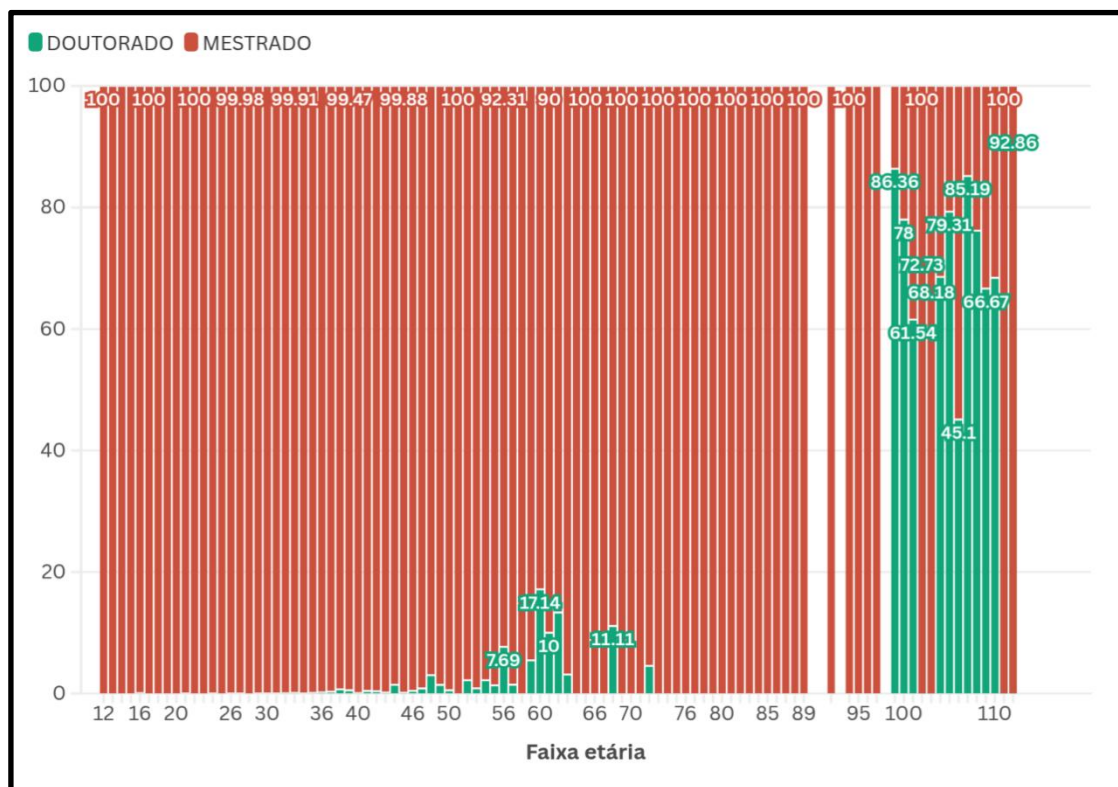
Em contrapartida, o doutorado acadêmico passa a apresentar maior representatividade a partir de 36 meses, com crescimento expressivo entre 45 e 54 meses — faixa que concentra mais de 95% das titulações no nível doutoral. Destacam-se os tempos de 48 meses (28.521 registros), 47 meses (22.864) e 45 meses (11.753), o que reflete o cumprimento da duração média de quatro anos, conforme as diretrizes da pós-graduação. Após 60 meses, os registros diminuem progressivamente, mas ainda há titulações com mais de 80 meses, possivelmente associadas a prorrogações autorizadas ou a circunstâncias específicas que levaram à extensão do período formativo.

A partir de 36 meses, o doutorado acadêmico passa a apresentar maior representatividade. No intervalo de 45 a 54 meses, concentra-se mais de 60% das titulações da modalidade, com destaque para 48 meses (28.521 registros, cerca de 12,7%), 47 meses (22.864 registros, 10,2%) e 45 meses (11.753, 5,2%). A partir de 60 meses, observa-se uma redução progressiva nos registros, mas ainda há cerca de 10% das titulações com tempo superior a cinco anos, indicando casos de prorrogação ou situações específicas.

Ressalta-se que os dados revelam padrões distintos e aderentes à regulamentação vigente. O mestrado acadêmico concentra a maioria das titulações entre 23 e 30 meses, com destaque para o pico em 24 meses (correspondente a dois anos), o que reflete a conclusão dentro do prazo estabelecido pela Resolução nº 5/1983. A partir de 36 meses, observa-se uma transição gradual, na qual o doutorado acadêmico passa a predominar, alcançando mais de 95% das titulações entre 45 e 54 meses, com pico em 48 meses (quatro anos), prazo convencionalmente associado à conclusão do doutorado acadêmico. Essa transição evidencia a clara separação entre os dois níveis de formação no que diz respeito à duração dos cursos. Assim, o gráfico confirma não apenas a aderência geral aos prazos regulamentares, mas também a existência de variações e flexibilidades no tempo de conclusão dos programas.

Dando sequência à análise do tempo de titulação, o Gráfico 44 explora esse indicador no contexto do mestrado e doutorado, ambos profissionais, possibilitando contrastes com os padrões observados nas modalidades acadêmicas.

Gráfico 44 – Comparação meses para titulação dos discentes doutorado e mestrado profissional; período 2004-2022



Fonte: Dados de pesquisa (2025).

Nos primeiros 24 meses, observa-se a predominância do mestrado profissional, com 15.851 registros (13,75%) no 24º mês e 12.623 registros (10,95%) no 23º mês, valores que refletem o padrão de duração esperado para essa modalidade. A partir do 30º mês, os números do mestrado profissional começam a declinar gradualmente: são 8.360 registros (7,25%) no 30º mês e 5.159 (4,48%) no 31º. O doutorado profissional, por sua vez, apresenta números bastante reduzidos nesse intervalo, com apenas 1 titulação em 24 meses e outra em 30 meses, representando frações percentuais irrelevantes.

A partir do 36º mês, começa-se a observar a presença mais frequente de titulações no doutorado profissional, embora ainda em proporções muito pequenas: 6 registros no 36º mês (0,01%) e crescimento moderado até o 48º mês, onde ocorrem 8 titulações (0,01%). Os valores do doutorado continuam reduzidos em comparação ao mestrado profissional, mantendo números inferiores a 10 registros por período.

A partir do 60º mês, os registros se tornam esparsos para ambas as modalidades, com titulações pontuais que podem indicar casos de prorrogação excepcional ou particularidades institucionais. O doutorado profissional apresenta alguns poucos registros acima de 100 meses, como 2 em 102 meses e 1 em 111 meses, representando apenas 0,002% do total analisado.

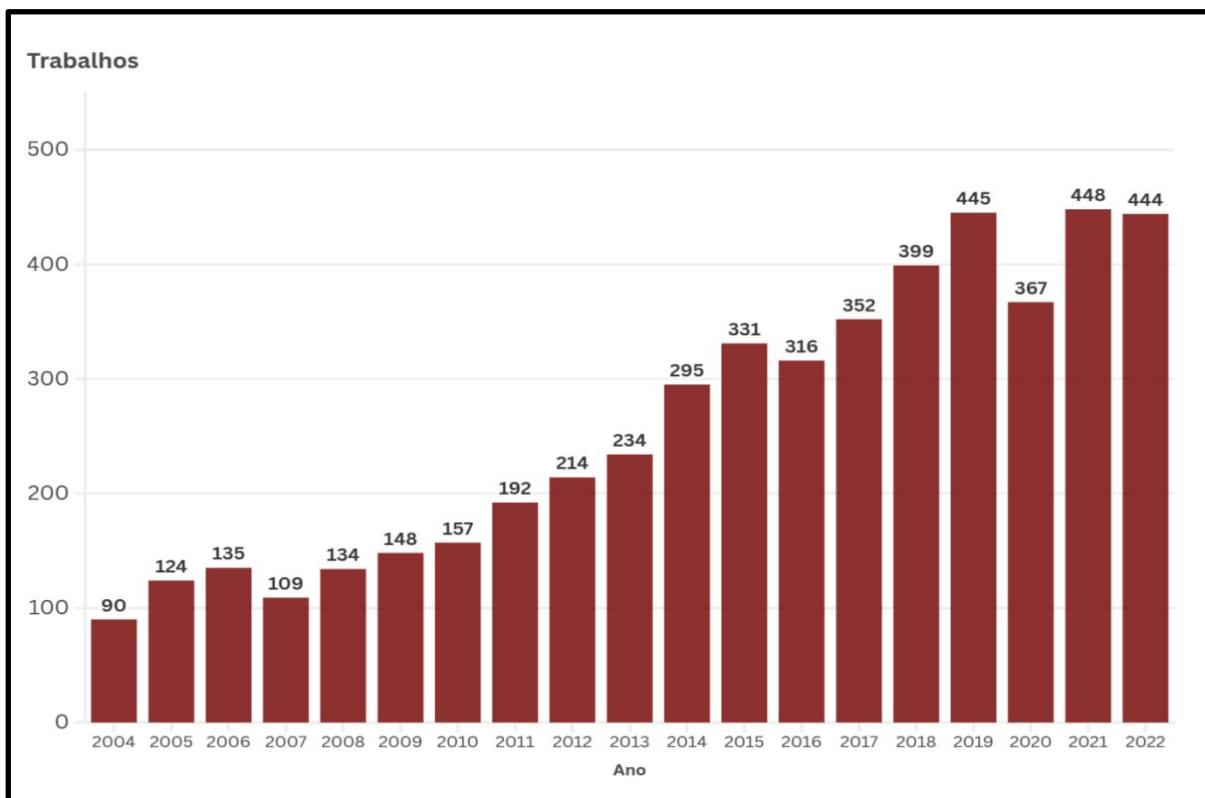
Com a finalização da etapa descritiva sobre o conjunto de dados dos discentes na pós-graduação brasileira, a análise agora se volta ao recorte dos dados vinculados à área da Ciência da Informação, ao conjunto de dados DF3_trabalhos_CI.

5.7.5 Resultados da etapa do Recorte dos dados da Ciência da Informação no DF3_trabalhos_CI

A aplicação do recorte no conjunto de dados DF3, com base no código de área do conhecimento 60700009, permite isolar os registros referentes exclusivamente à área de Ciência da Informação. Esse procedimento viabiliza a análise direcionada da produção acadêmica vinculada à pós-graduação *stricto sensu* nessa área, fornecendo subsídios para a compreensão de suas características institucionais. A partir desse recorte, é possível examinar variáveis como o ano de titulação, a instituição de ensino superior (IES), a região e a unidade federativa do programa, o grau de titulação, o nome do programa e o idioma dos trabalhos. A análise integrada dessas variáveis permite traçar um panorama comparativo da atuação da Ciência da Informação no sistema nacional de pós-graduação.

Iniciando a apresentação dos dados da área de Ciência da Informação, o Gráfico 45 exhibe a evolução do número de trabalhos defendidos entre 2004 e 2022, evidenciando variações e tendências na produção acadêmica ao longo do período.

Gráfico 45 – Número de trabalhos da CI defendidos de 2004 a 2022; período 2004-2022



Fonte: Dados de pesquisa (2025).

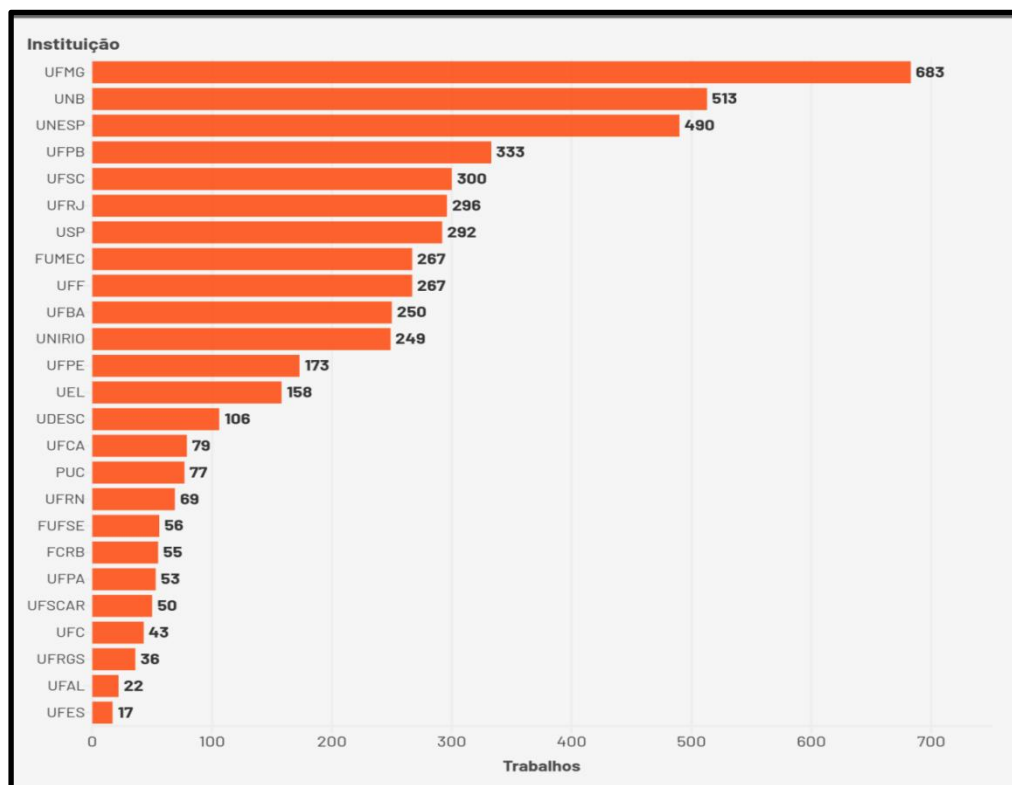
Observa-se que em 2004, foram defendidos 90 trabalhos (1,82% do total), número que cresceu nos anos seguintes: 124 em 2005 (2,51%) e 135 em 2006 (2,74%). Após uma leve queda em 2007 (109 trabalhos, 2,21%), os valores voltaram a subir, com 134 (2,72%) em 2008, ainda sem alcançar o valor de 2006, mas chegando a 148 trabalhos (3,00%) em 2009. A tendência de crescimento prosseguiu, com 157 registros em 2010 (3,18%), 192 em 2011 (3,89%) e 214 em 2012 (4,34%).

A expansão da área se intensificou a partir de 2013, com 234 trabalhos (4,74%), 295 em 2014 (5,98%), e 331 em 2015 (6,71%). De 2016 a 2019, os números seguiram em alta: 316 (6,40%) em 2016, 352 (7,13%) em 2017, 399 (8,09%) em 2018 e 445 em 2019 (9,02%).

Em 2020, ano marcado pelos impactos da pandemia da COVID-19, houve queda para 367 trabalhos (7,44%) — comportamento também observado em outras áreas, como apontado por Corrêa *et al.* (2022). A produção se recuperou em 2021, com 448 registros (9,08%), mantendo-se em patamar elevado em 2022, com 444 trabalhos (9,00%). Esse comportamento acompanha a tendência observada na pós-graduação brasileira como um todo, conforme evidenciado nos dados gerais apresentados anteriormente.

Dando continuidade à análise, o Gráfico 46 apresenta a distribuição dos trabalhos defendidos na área de Ciência da Informação entre 2004 e 2022, com destaque para as instituições de ensino superior (IES) responsáveis pela produção acadêmica no período.

Gráfico 46 – Distribuição dos trabalhos defendidos em CI IES; período 2004-2022



Fonte: Dados de pesquisa (2025).

Observa-se que a Universidade Federal de Minas Gerais (UFMG) lidera em número absoluto, com 683 registros (13,84%) no período analisado. Em seguida, destacam-se a Universidade de Brasília (UnB), com 513 trabalhos (10,40%), e a Universidade Estadual Paulista Júlio de Mesquita Filho (UNESP), com 490 registros (9,93%). Também apresentam alta representatividade a Universidade Federal da Paraíba (UFPB), com 333 trabalhos (6,75%), a Universidade Federal de Santa Catarina (UFSC), com 300 (6,08%), e a Universidade Federal do Rio de Janeiro (UFRJ-IBICT), com 296 (6,00%).

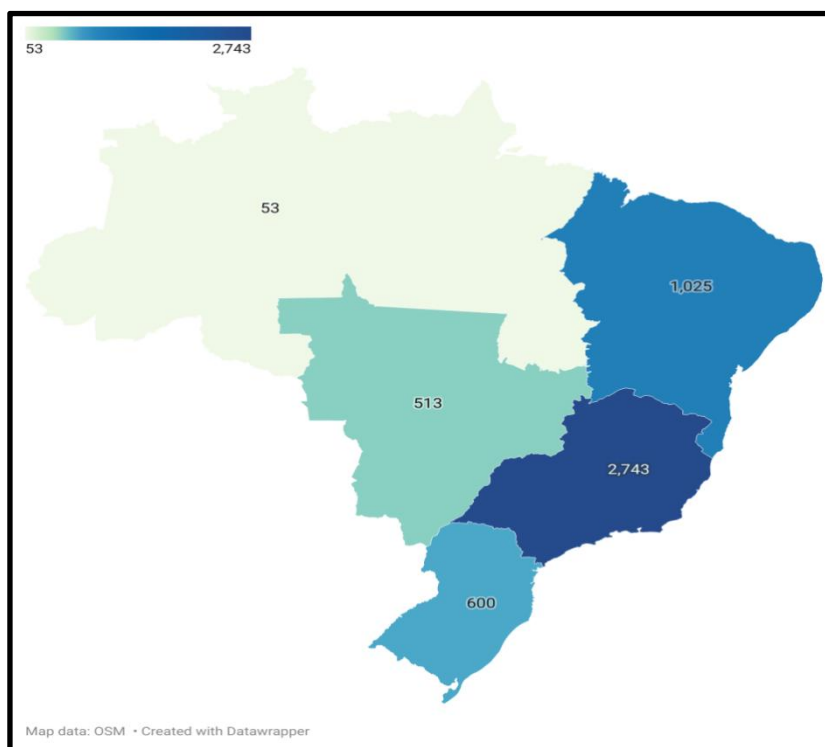
Outras instituições de destaque incluem a Universidade de São Paulo (USP), com 292 registros (5,92%), a Universidade FUMEC e a Universidade Federal Fluminense (UFF), ambas com 267 (5,41%), a Universidade Federal da Bahia (UFBA), com 250 (5,07%), e a Universidade Federal do Estado do Rio de Janeiro (UNIRIO), com 249 trabalhos (5,05%). A Universidade Federal de Pernambuco (UFPE) acumula 173 registros (3,51%), seguida pela Universidade Estadual de Londrina (UEL), com 158 (3,20%), e pela Universidade do Estado

de Santa Catarina (UDESC), com 106 (2,15%). Há instituições que apresentam menor participação relativa, como a Universidade Federal do Espírito Santo (UFES), com 17 registros (0,34%), e a Universidade Federal de Alagoas (UFAL), com 22 (0,45%), configurando um cenário de distribuição assimétrica da produção na área entre as diferentes instituições.

Destaca-se que a UFMG e UnB são os programas mais antigos, fundadores da área no Brasil. A UFSC, UFPB e UFRJ vieram na década de 1990, contribuindo para a expansão nacional da área. Por sua vez, USP, UFF, UFBA e FUMEC mostram que mesmo sendo programas mais recentes (pós-2000), conseguem alcançar alto volume de produção.

Dando continuidade à análise, o Gráfico 47 apresenta a distribuição das titulações na área de Ciência da Informação defendidas por região do país, permitindo observar os padrões espaciais da produção acadêmica no campo.

Gráfico 47 – Distribuição das titulações na CI por regiões do país; período 2004-2022



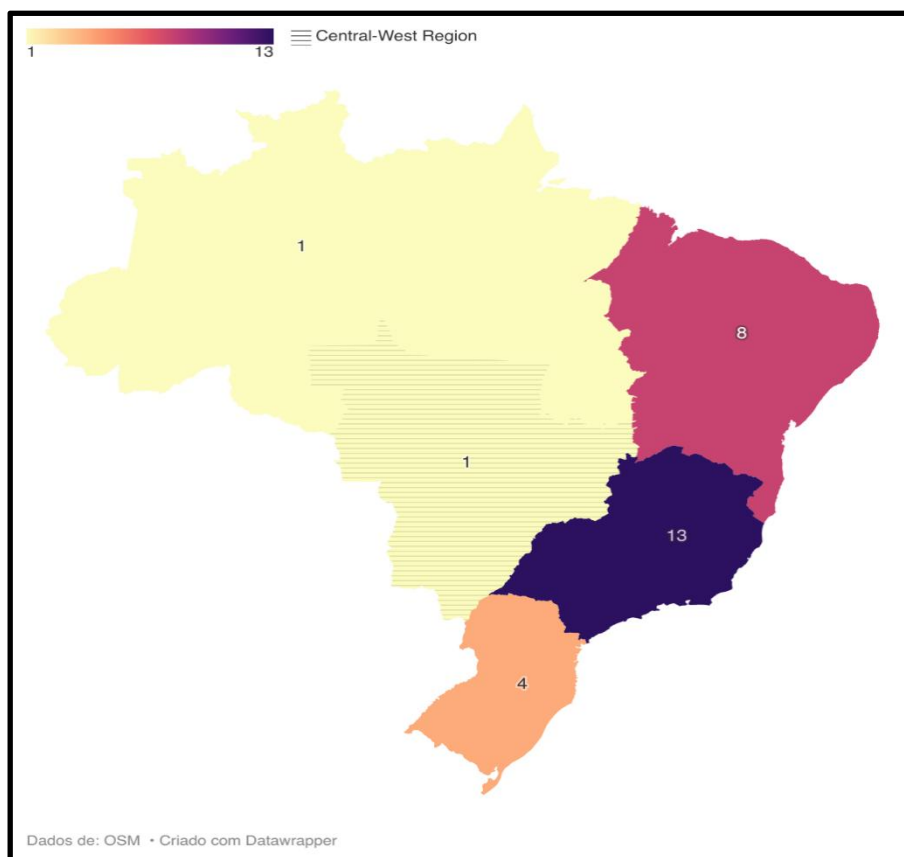
Fonte: Dados de pesquisa (2025).

Os dados evidenciam uma forte concentração na Região Sudeste, que responde por 2.743 trabalhos, o equivalente a 55,59% do total nacional. Em seguida, destaca-se a Região Nordeste, com 1.025 registros (20,77%), seguida pela Região Sul, com 600 trabalhos (12,16%). A Região Centro-Oeste apresenta 513 registros (10,40%), enquanto a Região Norte contabiliza apenas 53 trabalhos, representando 1,07% do total. Essa distribuição revela

disparidades regionais no desenvolvimento da pós-graduação em Ciência da Informação, possivelmente devido a iniciativas em infraestrutura acadêmica na área. Isso fica mais evidente com o segundo lugar em produção da Região Nordeste, apesar das desigualdades estruturais enfrentadas por essa região em outros aspectos (Ipea, 2021).

Aprofundando essa análise, o Gráfico 48 apresenta a distribuição das IES (onde ficam os programas de pós-graduação) por região na área da Ciência da Informação, complementando os padrões identificados no Gráfico 47.

Gráfico 48 – Distribuição das IES por região; período 2004-2022



Fonte: Dados de pesquisa (2025).

A distribuição das Instituições de Ensino Superior (IES) com programas de pós-graduação em Ciência da Informação no Brasil, no período de 2004 a 2022, revela forte concentração regional. O Sudeste lidera com 13 IES, representando 48,2% do total nacional, com distribuição bastante equilibrada entre três estados. Minas Gerais, com 4 IES (UFMG, UFOP, FUMEC, UFU), é o estado com maior concentração de programas na região. São Paulo vem logo em seguida, com 3 IES (USP, UNESP, PUC-SP). O Rio de Janeiro também conta com 3 instituições (UFRJ-IBICT, UFF, UNIRIO), enquanto o Espírito Santo possui apenas 1 (UFES), com menor expressividade regional.

Em seguida, o Nordeste aparece com 8 programas (29,6%), indicando avanço importante na interiorização e diversificação da pós-graduação no campo. A região é liderada pela Paraíba, com 2 IES (UFPB e UFCG), seguida pela Bahia (UFBA) e Pernambuco (UFPE), com 1 IES cada. Também há 1 IES no Ceará (UFCA), Rio Grande do Norte (UFRN) e Sergipe (UFS). Apesar da diversidade de estados com programas, o número de instituições por estado sugere espaço para expansão.

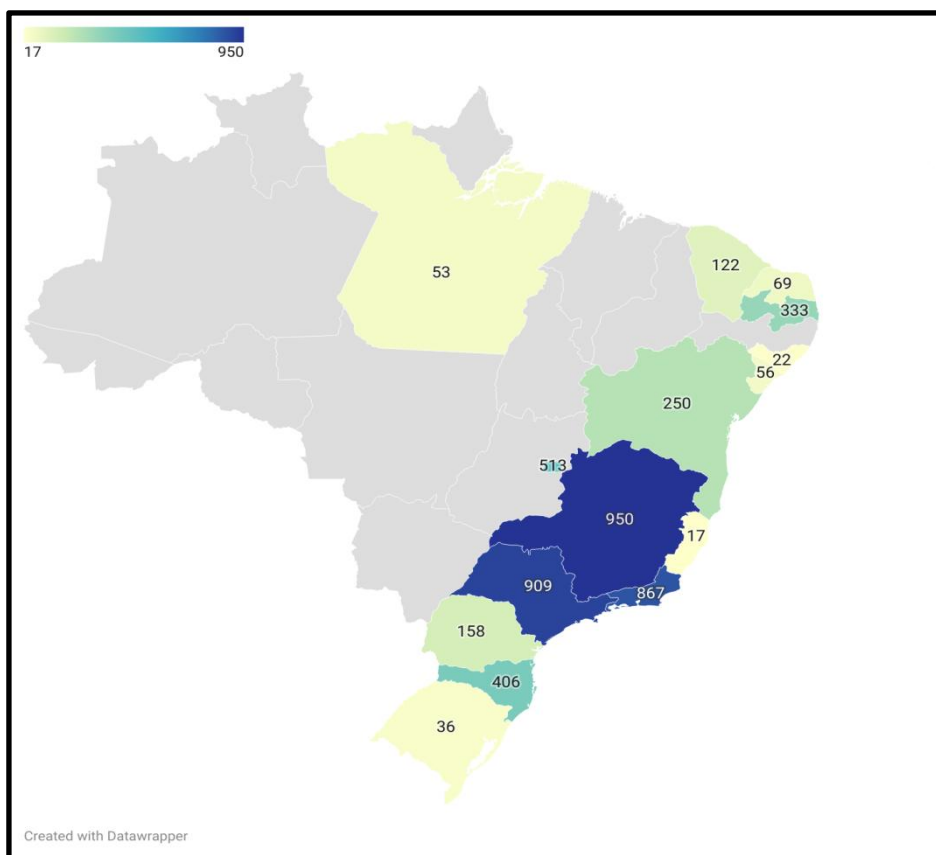
O Sul conta com 4 IES (14,8%), sendo Santa Catarina o estado com maior número de programas da região, com 2 IES (UFSC e UDESC). Paraná conta com 1 IES (UEL) e o Rio Grande do Sul também com 1 (UFRGS), refletindo uma distribuição mais dispersa, porém, ainda concentrada nos estados mais populosos.

O Centro-Oeste, com 1 IES (3,7%), tem toda a produção da região concentrada na Universidade de Brasília (UnB), no Distrito Federal, o que revela ausência de programas nos demais estados do Centro-Oeste (GO, MT e MS). Isso representa uma lacuna na interiorização da região.

O Norte, também com 1 IES (3,7%), conta apenas com o Pará com programa ativo na região, por meio da UFPA. Os demais estados da região (AC, AM, AP, RO, RR e TO) não apresentam programas *stricto sensu* em Ciência da Informação, evidenciando também espaço para expansão.

Dando prosseguimento à análise, o Gráfico 49 explora a concentração das titulações na área de Ciência da Informação por Unidade da Federação (UF), destacando as dinâmicas estaduais de produção acadêmica.

Gráfico 49 – Quantidade de titulações na CI por UF; período 2004-2022



Fonte: Dados de pesquisa (2025).

Os dados revelam forte concentração nas UFs do Sudeste. Minas Gerais lidera em número absoluto, com 950 trabalhos defendidos, o que corresponde a 19,25% do total nacional. Em seguida, aparecem São Paulo, com 909 registros (18,42%), e o Rio de Janeiro, com 867 trabalhos (17,57%). O Espírito Santo apresenta participação mais modesta, com 17 registros (0,34%).

Na Região Sul, Santa Catarina contabiliza 406 titulações (8,23%) e o Paraná, 158 (3,20%). No Nordeste, a Paraíba se sobressai com 333 registros (6,75%), seguida pela Bahia, com 250 (5,07%), e Pernambuco, com 173 (3,51%). Outros estados da região também contribuem para a produção, como o Ceará, com 122 trabalhos (2,47%), o Rio Grande do Norte, com 69 (1,40%), e Sergipe, com 56 registros (1,13%).

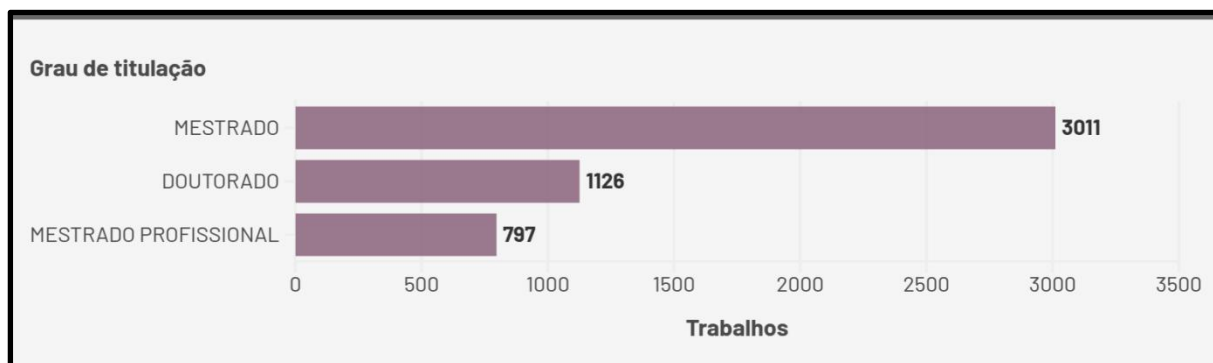
Na Região Norte, há apenas o estado do Pará, com 53 titulações (1,07%). Por fim, o Centro-Oeste, também representado pelo Distrito Federal, com 513 registros (10,40%), que centraliza toda a produção da região.

Cabe destacar que diversas UFs não apresentam nenhum trabalho no período analisado, incluindo Acre, Amapá, Amazonas, Goiás, Maranhão, Mato Grosso, Mato Grosso do Sul,

Piauí, Rondônia, Roraima e Tocantins, evidenciando importantes lacunas regionais na oferta de programas de pós-graduação em Ciência da Informação.

O Gráfico 50 amplia essa análise ao apresentar a distribuição dos trabalhos defendidos na área de Ciência da Informação segundo o grau de titulação, permitindo observar a prevalência entre mestrado e doutorado no período considerado.

Gráfico 50 – Quantidade de trabalhos da CI por grau de titulação; período 2004-2022



Fonte: Dados de pesquisa (2025).

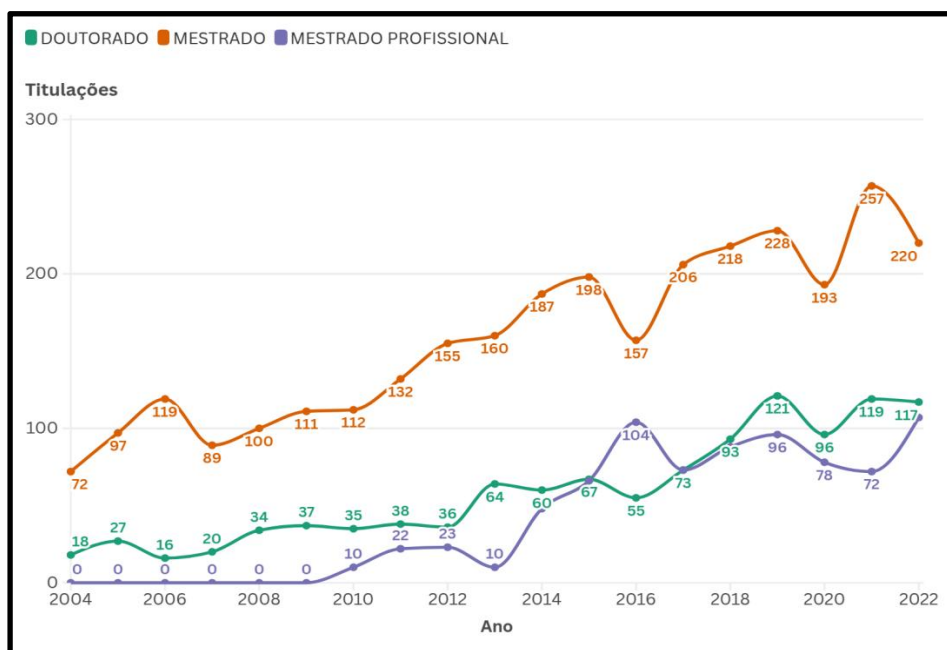
Os dados indicam que as dissertações de mestrado acadêmico representam a maior parte das produções na área, com 3.011 trabalhos, o equivalente a 61,03% do total. O mestrado profissional responde por 797 registros (16,15%), enquanto o doutorado acadêmico contabiliza 1.126 trabalhos, representando 22,82% das titulações analisadas.

Ao agrupar os dois tipos de mestrado (acadêmico e profissional), observa-se que somam 3.808 registros, que correspondem a 77,18% das defesas. Esse dado evidencia a centralidade do mestrado no sistema de pós-graduação em Ciência da Informação. Essa predominância também se explica pela própria estrutura dos programas: o mestrado, com duração média de dois anos, tende a gerar o dobro de titulações em comparação ao doutorado, cuja duração média é de quatro anos. Desse modo, a diferença temporal entre os cursos naturalmente impacta o volume de produções acadêmicas.

Até 2022, não foram identificadas titulações de doutorado profissional na área da Ciência da Informação, o que reflete tanto a recente regulamentação dessa modalidade quanto a sua ainda limitada adoção nos programas da área.

O Gráfico 51 detalha a evolução anual dos trabalhos defendidos na área da Ciência da Informação entre 2004 e 2022, segmentados por grau de titulação. A visualização permite observar o comportamento dinâmico da produção acadêmica ao longo do tempo, destacando tendências específicas entre mestrado acadêmico, mestrado profissional e doutorado acadêmico.

Gráfico 51 – Números de trabalho da CI por grau em cada ano; período 2004-2022



Fonte: Dados de pesquisa (2025).

O mestrado acadêmico lidera em volume absoluto de titulações e apresenta uma trajetória de crescimento consistente ao longo do período. Em 2004, foram registrados 72 trabalhos, número que salta para 119 já em 2006. Após um período de estabilidade entre 2007 e 2009, observa-se novo avanço a partir de 2010, com destaques em 2014 (187), 2015 (198) e 2019 (228). O pico da série foi registrado em 2021, com 257 dissertações defendidas, seguido por uma leve retração para 220 em 2022.

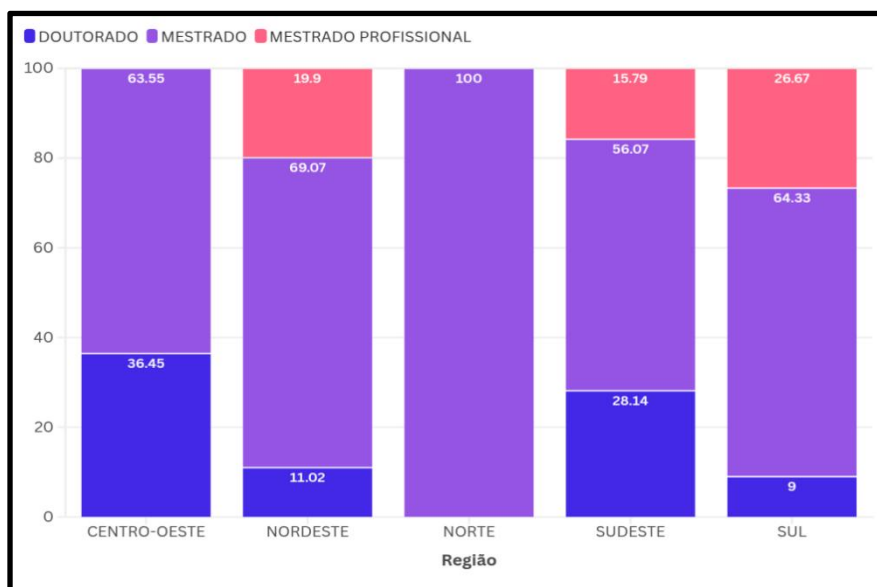
O mestrado profissional, regulamentado mais recentemente, inicia sua trajetória em 2010 com 10 trabalhos. A partir de 2013, observa-se crescimento expressivo, passando de 10 para 48 trabalhos em 2014, atingindo 66 registros em 2015 e 104 em 2016. Apesar de pequenas oscilações, a trajetória é ascendente, pois em 2019 o número chegou a 96 trabalhos e 2020 chegou a 78 trabalhos, tendo se recuperado com novo recorde de 107 trabalhos em 2022. Trata-se do grau com maior crescimento relativo no período.

O doutorado acadêmico, por sua vez, apresenta avanço mais gradual. Em 2004, foram defendidas 18 teses, com crescimento moderado até atingir 121 em 2019, seu ponto mais alto. O maior salto ocorreu entre 2018 e 2019, passando de 93 para 121 defesas. Nos anos seguintes, os números se mantiveram estáveis: 119 em 2021 e 117 em 2022.

A análise revela que, embora o mestrado acadêmico mantenha a predominância em volume, o mestrado profissional desponta como modalidade em expansão, enquanto o doutorado avança de forma constante.

O Gráfico 52 aprofunda a análise ao apresentar a distribuição percentual dos graus acadêmicos na área da Ciência da Informação por região. A visualização permite compreender como cada modalidade — mestrado acadêmico, mestrado profissional e doutorado acadêmico — se configura nos diferentes contextos regionais do país.

Gráfico 52 – Distribuição percentual dos graus acadêmicos da CI por região; período 2004-2022



Fonte: Dados de pesquisa (2025).

Na Região Centro-Oeste, observa-se uma divisão exclusiva entre mestrado acadêmico e doutorado acadêmico. O mestrado acadêmico responde por 63,55% dos trabalhos (326), enquanto o doutorado representa 36,45% (187), tendo a UnB como única IES da região com programa ativo na área.

No Nordeste, o mestrado acadêmico é predominante, com 69,07% das titulações (708 trabalhos), seguido pelo mestrado profissional, que representa 19,90% (204), e o doutorado acadêmico, com 11,02% (113). Essa composição indica maior diversidade de modalidades e boa adesão à formação profissional na região.

A Região Norte concentra 100% das suas titulações no mestrado acadêmico (53 trabalhos), sem registros de doutorado acadêmico ou mestrado profissional, revelando limitações na oferta dessas modalidades e oportunidades formativas na área.

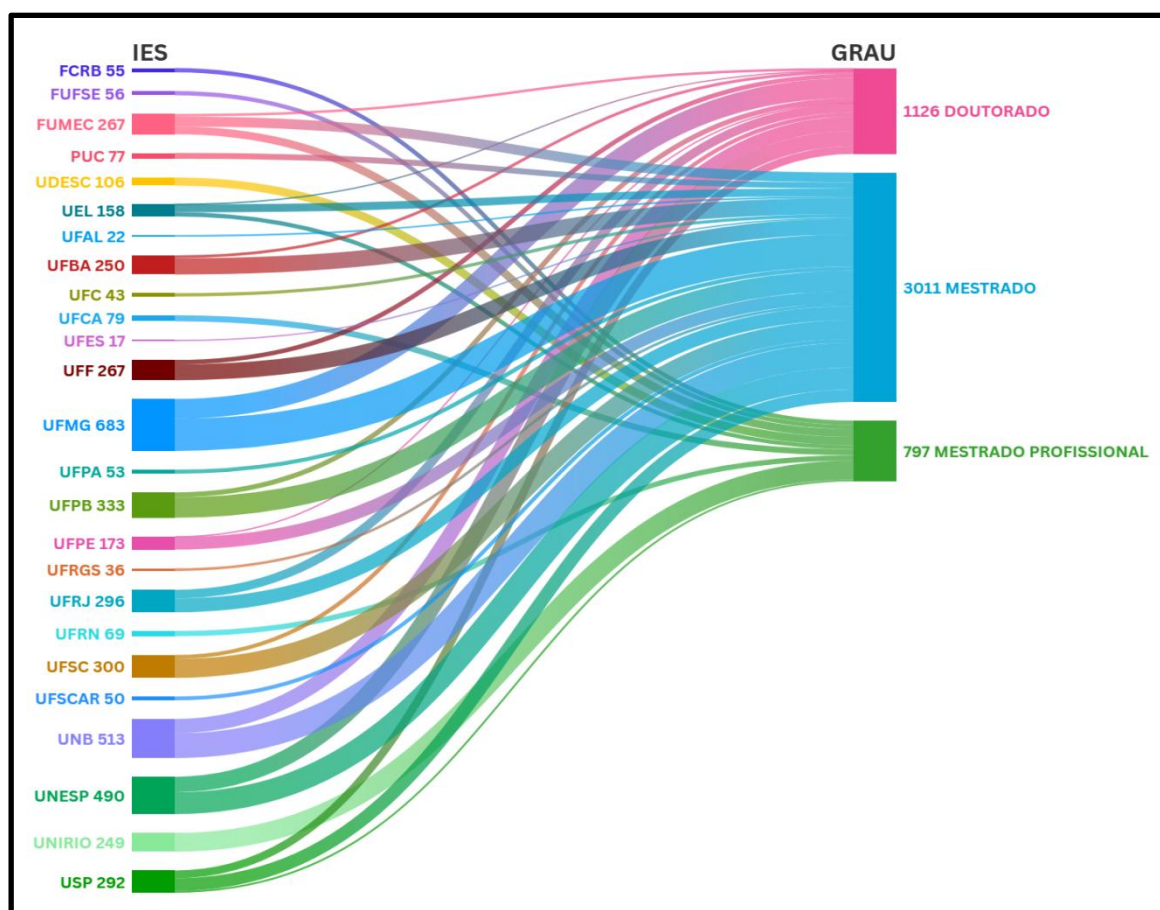
No Sudeste, o mestrado acadêmico também lidera, com 56,07% das titulações (1.538 trabalhos). O doutorado acadêmico responde por 28,14% (772), enquanto o mestrado profissional representa 15,79% (433). Trata-se da região com maior volume absoluto de trabalhos.

No Sul, o mestrado acadêmico também é majoritário, com 64,33% dos registros (386 trabalhos). O mestrado profissional assume papel relevante, com 26,67% (160), enquanto o doutorado acadêmico apresenta participação mais modesta, com 9,00% (54).

Esses dados indicam que, embora o mestrado acadêmico seja predominante em todas as regiões, como esperado, o mestrado profissional tem ganhado destaque especialmente no Sul e Nordeste, enquanto o doutorado acadêmico apresenta maior peso relativo no Sudeste e no Centro-Oeste.

O Gráfico 53 complementa essa análise ao apresentar a distribuição das Instituições de Ensino Superior (IES) por grau acadêmico na área da Ciência da Informação. Do lado esquerdo, são listadas as instituições; do lado direito, os graus de titulação oferecidos — mestrado, mestrado profissional e doutorado acadêmico.

Gráfico 53 – Distribuição das IES por grau acadêmico; período 2004-2022



Fonte: Dados de pesquisa (2025).

Observa-se que diversas instituições ofertam mais de um grau de titulação. A Universidade Federal de Minas Gerais (UFMG), com 423 mestrados acadêmicos e 260 doutorados acadêmicos, a Universidade de Brasília (UnB) reúne 326 mestrados acadêmicos e

187 doutorados acadêmicos, e a Universidade Estadual Paulista (UNESP) se destaca com 288 mestrados acadêmicos e 202 doutorados acadêmicos.

Outras instituições com expressivo volume de titulações incluem a Universidade Federal do Rio de Janeiro (UFRJ-IBICT), com 181 mestrados acadêmicos e 115 doutorados acadêmicos, e a Universidade de São Paulo (USP), que apresenta 165 defesas de mestrado acadêmico, 25 de mestrado profissional e 102 de doutorado acadêmico, uma das instituições que apresentam os três níveis de titulação.

A Universidade FUMEC é outro exemplo de instituição que oferece os três níveis de formação, com 130 mestrados acadêmicos, 104 mestrados profissionais e 33 doutorados acadêmicos. Entre as que se destacam no mestrado profissional, encontram-se a Universidade Federal do Estado do Rio de Janeiro (UNIRIO), com 249 registros, a Universidade Federal do Cariri (UFCA), com 79, e a Universidade Estadual de Londrina (UEL), com 54.

Algumas instituições se concentram exclusivamente em um único grau. A UDESC, por exemplo, apresenta 106 registros de mestrado profissional, e a Fundação Casa de Rui Barbosa (FCRB) tem 55, ambos sem doutorado. Já a UFPA tem apenas mestrado acadêmico (53 registros), e a UFRGS, 36 mestrados acadêmicos, sem registros de doutorado na área.

Os 31 programas vinculados à área da Ciência da Informação, identificados nos dados analisados e apresentados na Tabela 36, estão distribuídos em 25 instituições de ensino superior. A maioria adota a denominação Ciência da Informação, mas também há programas com nomes que refletem enfoques específicos, como: Gestão da Informação, Memória e Acervos, Gestão & Organização do Conhecimento, Biblioteconomia, Gestão de Documentos e Arquivos, Tecnologia da Informação e Comunicação e Gestão do Conhecimento, Sistemas de Informação e Gestão do Conhecimento. Destaca-se ainda o programa da UFRJ, identificado como UFRJ-IBICT, em referência ao Instituto Brasileiro de Informação em Ciência e Tecnologia, que compõe sua estrutura. Destaca-se a particularidade na denominação do programa da UFRJ, identificado como UFRJ-IBICT (em referência ao Instituto Brasileiro de Informação em Ciência e Tecnologia), que usa o nome da universidade seguido da sigla IBICT.

Tabela 36 – Programas de pós-graduação stricto sensu em Ciência da Informação; período 2004-2022

Índice	Código programa	Nome do programa	IES	Trabalhos
1	33301000001P0	MEMÓRIA E ACERVOS	FUNDACAO CASA DE RUI BARBOSA	55
2	27001016175P0	CIÊNCIA DA INFORMAÇÃO	FUNDAÇÃO UNIVERSIDADE	56

Índice	Código programa	Nome do programa	IES	Trabalhos
			FEDERAL DE SERGIPE	
3	33006016003P4	CIÊNCIA DA INFORMAÇÃO	PONTIFÍCIA UNIVERSIDADE CATÓLICA	77
4	53001010018P0	CIÊNCIA DA INFORMAÇÃO	UNIVERSIDADE DE BRASÍLIA	513
5	33002010195P5	CIÊNCIA DA INFORMAÇÃO	UNIVERSIDADE DE SÃO PAULO	267
6	33002010248P1	GESTÃO DA INFORMAÇÃO	UNIVERSIDADE DE SÃO PAULO	25
7	41002016025P5	GESTÃO DA INFORMAÇÃO	UNIVERSIDADE DO ESTADO DE SANTA CATARINA	106
8	40002012035P8	GESTÃO DA INFORMAÇÃO	UNIVERSIDADE ESTADUAL DE LONDRINA	54
9	40002012048P2	CIÊNCIA DA INFORMAÇÃO	UNIVERSIDADE ESTADUAL DE LONDRINA	104
10	33004110043P4	CIÊNCIA DA INFORMAÇÃO	UNIVERSIDADE ESTADUAL PAULISTA JÚLIO DE MESQUITA FILHO	490
11	28001010041P0	CIÊNCIA DA INFORMAÇÃO	UNIVERSIDADE FEDERAL DA BAHIA	250
12	24001015049P7	CIÊNCIA DA INFORMAÇÃO	UNIVERSIDADE FEDERAL DA PARAIBA	333
13	26001012171P2	CIÊNCIA DA INFORMAÇÃO	UNIVERSIDADE FEDERAL DE ALAGOAS	22
14	32001010028P2	CIÊNCIA DA INFORMAÇÃO	UNIVERSIDADE FEDERAL DE MINAS GERAIS	540
15	32001010172P6	GESTÃO & ORGANIZAÇÃO DO CONHECIMENTO	UNIVERSIDADE FEDERAL DE MINAS GERAIS	143
16	25001019077P3	CIÊNCIA DA INFORMAÇÃO	UNIVERSIDADE FEDERAL DE PERNAMBUCO	173
17	41001010052P0	CIÊNCIA DA INFORMAÇÃO	UNIVERSIDADE FEDERAL DE SANTA CATARINA	300
18	33001014052P3	CIÊNCIA DA INFORMAÇÃO	UNIVERSIDADE FEDERAL DE SÃO CARLOS	50
19	22033017002P3	BIBLIOTECONOMIA	UNIVERSIDADE FEDERAL DO CARIRI	79
20	22001018085P8	CIÊNCIA DA INFORMAÇÃO	UNIVERSIDADE FEDERAL DO CEARÁ	43
21	30001013108P0	CIÊNCIA DA INFORMAÇÃO	UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO	17
22	31021018018P8	BIBLIOTECONOMIA	UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO	164
23	31021018019P4	GESTÃO DE DOCUMENTOS E ARQUIVOS	UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO	85
24	15001016158P5	CIÊNCIA DA INFORMAÇÃO	UNIVERSIDADE FEDERAL DO PARÁ	53
25	31001017138P0	CIÊNCIA DA INFORMAÇÃO - UFRJ – IBICT	UNIVERSIDADE FEDERAL DO RIO DE JANEIRO	296
26	23001011080P9	CIÊNCIA DA INFORMAÇÃO	UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE	69
27	42001013176P0	CIÊNCIA DA INFORMAÇÃO	UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL	36
28	31003010056P0	CIÊNCIA DA INFORMAÇÃO	UNIVERSIDADE FEDERAL FLUMINENSE	89
29	31003010063P6	CIÊNCIA DA INFORMAÇÃO	UNIVERSIDADE FEDERAL FLUMINENSE	178
30	32053010003P1	TECNOLOGIA DA INFORMAÇÃO E	UNIVERSIDADE FUMEC	265

Índice	Código programa	Nome do programa	IES	Trabalhos
		COMUNICAÇÃO E GESTÃO DO CONHECIMENTO		
31	32053010006P0	SISTEMAS DE INFORMAÇÃO E GESTÃO DO CONHECIMENTO	UNIVERSIDADE FUMEC	2

Fonte: Dados de pesquisa (2025).

A Tabela 36 reúne os programas, indicando sua situação atual no Sistema Nacional de Pós-Graduação (SNPG). Os programas destacados com células em vermelho correspondem àqueles desativados no sistema da Capes.

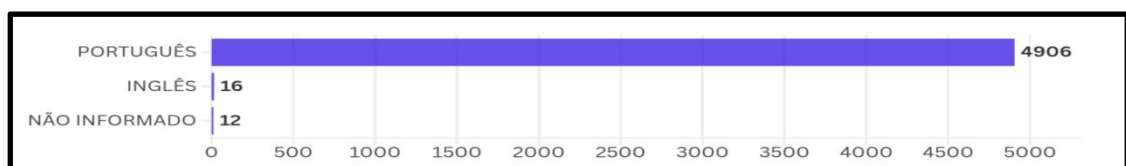
Entre os cursos extintos, há o Programa de Ciência da Informação da Pontifícia Universidade Católica de Campinas (PUC-Campinas), descredenciado após obter conceito inferior a 3 na Avaliação Trienal de 2007. Situação semelhante ocorreu com o Programa de Gestão da Informação da Universidade Estadual de Londrina (UEL), cuja continuidade foi inviabilizada por não alcançar a nota mínima exigida pelo SNPG, tendo recebido nota 2 na avaliação na Avaliação Quadrienal 2017-2020.

O programa de Ciência da Informação da Universidade Federal Fluminense (UFF), por sua vez, foi descontinuado após o encerramento do Convênio UFF-IBICT, ocorrido em outubro de 2008, o qual sustentava sua operacionalização.

O último caso é o da Universidade FUMEC, cujo programa Sistemas de Informação e Gestão do Conhecimento passou por um processo de fusão entre as versões de mestrado (32053010003P1) e doutorado (32053010006P0), anteriormente tratadas como programas distintos. A partir dessa reestruturação, foi criado um único programa ativo, com ênfase profissional em ambos os níveis (mestrado e doutorado), mantendo-se registrado na base da Capes. Conforme registrado na Plataforma Sucupira da CAPES, a fusão desses programas ocorreu antes da publicação da Portaria 256/2018 (Brasil, 2018), que estabeleceu diretrizes para a fusão, desmembramento e migração de programas de pós-graduação *stricto sensu* no país.

Dando continuidade à análise, o Gráfico 54 apresenta a distribuição dos idiomas em que foram redigidos os trabalhos da área da Ciência da Informação no período analisado.

Gráfico 54 – Idioma dos trabalhos da CI; período 2004-2022

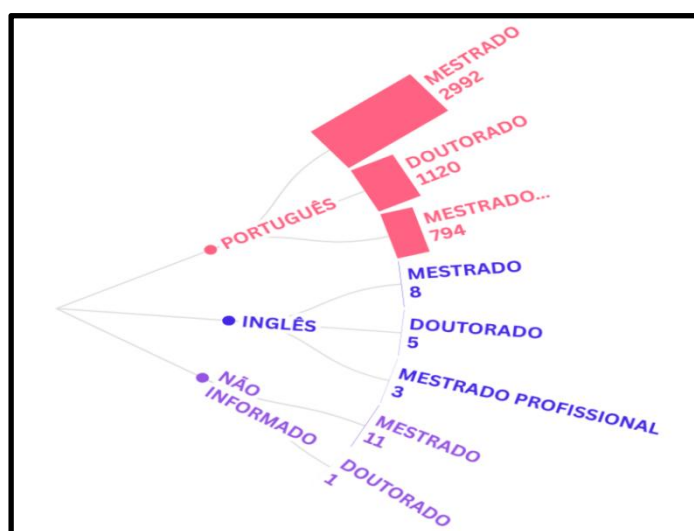


Fonte: Dados de pesquisa (2025).

A predominância do idioma português é marcante, como esperado, com 4.906 trabalhos, o que representa 99,43% do total. O inglês, embora seja o idioma predominante da comunicação científica internacional (Montgomery, 2013), aparece com uma participação residual, contabilizando 16 trabalhos (0,32%), o que pode representar um desafio à internacionalização da área. Além disso, há 12 registros (0,24%) classificados como “não informado”, o que pode indicar ausência de metadados completos, inconsistências no preenchimento ou indisponibilidade do texto integral nos sistemas consultados.

Complementado essa análise, o Gráfico 55 sintetiza a relação entre o grau acadêmico e o idioma dos trabalhos defendidos na área da Ciência da Informação.

Gráfico 55 – Grau acadêmico dos trabalhos da CI por idioma; período 2004-2022



Fonte: Dados de pesquisa (2025).

A predominância do português se mantém em todos os níveis de formação, como esperado. Dos trabalhos redigidos nesse idioma, 2.992 (61%) correspondem ao mestrado acadêmico, 1.120 (22,82%) ao doutorado acadêmico e 794 (16,18%) ao mestrado profissional, totalizando 4.906 registros. Já os trabalhos em inglês, estão distribuídos entre 8 dissertações de mestrado acadêmico, 3 de mestrado profissional e 5 teses de doutorado, somando 16 registros. Por fim, entre os trabalhos classificados como “não informado” quanto ao idioma, identificam-se 11 dissertações de mestrado e 1 tese de doutorado.

Passando agora à análise dos títulos dos trabalhos acadêmicos (teses e dissertações), o Gráfico 56 apresenta um *heatmap* com as 20 palavras mais frequentes, evidenciando as coocorrências mais recorrentes entre elas. A visualização permite observar as associações léxicas mais consolidadas na produção da área da Ciência da Informação.

Gráfico 56 – Coocorrência das 20 palavras mais utilizadas nos títulos dos trabalhos da CI; período 2004-2022

informação	0	350	329	262	96	74	85	56	71	84	117	144	139	161	93	160	232	109	155	144
ciência	350	0	66	73	24	14	14	4	18	46	16	66	26	10	14	24	21	29	44	90
estudo	329	66	0	42	61	7	36	31	23	42	33	54	50	59	63	94	99	74	76	67
análise	262	73	42	0	53	21	42	29	26	57	42	63	42	35	27	46	57	39	73	90
informacional	96	24	61	53	0	12	20	7	4	14	9	13	37	13	31	28	25	37	9	15
proposta	74	14	7	21	12	0	7	16	10	4	16	3	25	27	23	42	35	32	21	18
rede	85	14	36	42	20	7	0	5	4	10	7	13	18	9	5	35	10	12	25	24
arquivo	56	4	31	29	7	16	5	0	20	6	12	19	17	12	14	10	25	22	3	4
memória	71	18	23	26	4	10	4	20	0	9	9	9	13	4	9	13	8	16	8	9
produção	84	46	42	57	14	4	10	6	9	0	11	39	11	6	20	6	15	15	40	127
organização	117	16	33	42	9	16	7	12	9	11	0	11	14	29	8	11	13	5	98	12
brasil	144	66	54	63	13	3	13	19	9	39	11	0	19	9	5	32	24	6	27	61
digital	139	26	50	42	37	25	18	17	13	11	14	19	0	12	18	71	22	26	11	20
sistema	161	10	59	35	13	27	9	12	4	6	29	9	12	0	31	51	46	28	36	12
universidade	93	14	63	27	31	23	5	14	9	20	8	5	18	31	0	60	40	171	21	28
biblioteca	160	24	94	46	28	42	35	10	13	6	11	32	71	51	60	0	43	70	24	16
gestão	232	21	99	57	25	35	10	25	8	15	13	24	22	46	40	43	0	55	141	22
federal	109	29	74	39	37	32	12	22	16	15	5	6	26	28	171	70	55	0	19	24
conhecimento	155	44	76	73	9	21	25	3	8	40	98	27	11	36	21	24	141	19	0	38
científico	144	90	67	90	15	18	24	4	9	127	12	61	20	12	28	16	22	24	38	0
informação		ciência	estudo	análise	informacional	proposta	rede	arquivo	memória	produção	organização	brasil	digital	sistema	universidade	biblioteca	gestão	federal	conhecimento	científico

Fonte: Dados de pesquisa (2025).

Observa-se que a associação mais frequente é entre as palavras “informação” e “ciência”, que aparecem juntas em 350 títulos, representando 7,09% do total de trabalhos analisados. Logo em seguida, destaca-se a combinação “informação” e “estudo”, com 329 coocorrências (6,67%), e “informação” e “análise”, com 262 registros (5,31%). Essas combinações revelam uma forte centralidade do termo “informação”, que aparece associado a conceitos estruturantes da área.

Também se destacam os pares “informação” e “gestão” (232 coocorrências, 4,70%) e “informação” e “conhecimento” (155 registros, 3,14%), que indicam o foco na relação entre os processos informacionais e a produção, organização e uso do conhecimento. Já o par “universidade” e “federal” aparece em 171 títulos (3,47%), possivelmente refletindo a ambientação do estudo ou a identidade institucional dos programas de pós-graduação.

Outros pares de destaque incluem: “informação” e “sistema” – 161 registros (3,26%); “informação” e “biblioteca” – 160 registros (3,24%); “informação” e “Brasil” – 144 registros (2,92%); “informação” e “científico” – 144 registros (2,92%); “gestão” e “conhecimento” – 141 registros (2,86%); e “informação” e “digital” – 139 registros (2,82%).

Essas relações destacam os principais eixos temáticos que permeiam a produção científica na área, como a interface entre informação, conhecimento, ciência e gestão, bem como as abordagens tecnológicas e ambientais presentes nos títulos. O *heatmap* contribui para identificar padrões semânticos e a recorrência de temas centrais no campo da Ciência da Informação.

Dando sequência à análise lexical da produção acadêmica, o Gráfico 57 apresenta um *heatmap* que representa a distribuição das 20 palavras mais utilizadas nos títulos das teses e dissertações no período analisado. Essa visualização permite observar a frequência absoluta dos termos ao longo do tempo, possibilitando a identificação de tendências temáticas e a evolução da terminologia empregada na titulação dos trabalhos no período analisado.

Gráfico 57 – TOP 20 palavras mais frequentes nos títulos dos trabalhos da CI por ano; período 2004-2022

Palavras	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
informação -	66	77	89	57	80	80	75	121	106	111	146	145	133	137	130	173	141	198	160
ciência -	7	9	11	9	10	19	18	15	18	19	25	24	29	27	26	57	37	46	47
estudo -	18	26	20	17	21	20	38	28	37	37	39	66	52	53	56	62	46	59	60
análise -	11	24	12	12	19	15	21	22	26	32	47	50	39	44	62	71	60	55	59
informacional -	9	11	10	8	15	15	17	28	22	22	19	21	23	26	28	36	27	35	32
proposta -	1	7	6	2	9	1	4	3	9	10	10	10	20	20	33	25	20	23	40
rede -	7	2	4	5	4	7	7	8	15	7	16	21	13	19	14	19	15	23	14
arquivo -	1	5	3	4	4	2	5	8	10	11	15	9	15	30	22	40	22	40	28
memória -	1	2	1	4	1	4	6	7	16	14	8	17	20	18	17	20	23	28	26
produção -	5	4	10	5	2	9	11	11	12	8	15	23	17	22	12	22	21	17	12
organização -	3	5	4	4	6	6	10	10	17	15	18	17	8	20	20	19	15	18	20
brasil -	8	11	4	9	9	12	10	16	13	13	18	18	24	19	28	28	32	27	32
digital -	2	12	6	8	10	14	22	16	19	16	17	24	25	24	35	34	39	35	43
sistema -	14	11	8	8	11	8	11	10	8	12	21	21	18	18	24	27	18	24	26
universidade -	0	3	7	3	1	7	11	6	25	7	16	15	21	17	20	22	32	21	29
biblioteca -	9	7	11	7	19	14	18	17	30	31	42	51	48	43	65	61	49	49	56
gestão -	5	15	21	16	7	15	17	17	15	23	20	29	35	52	41	63	40	40	60
federal -	3	4	2	2	4	6	11	3	11	13	15	26	26	21	25	27	39	37	34
conhecimento -	8	14	20	13	11	23	18	15	20	17	29	31	25	41	30	36	30	27	40
científico -	6	10	13	6	9	16	18	19	24	19	23	43	25	30	26	39	39	30	33

Fonte: Dados de pesquisa (2025).

Observa-se que as palavras com maior crescimento absoluto foram três: "informação", que passou de 66 ocorrências em 2004 para 160 em 2022, representando um aumento de 142,42%; "gestão", que saltou de 5 para 60 registros, totalizando 1100% de crescimento; e "digital", que subiu de 2 para 43, com um expressivo aumento de 2050%. Esses dados refletem a centralidade dessas temáticas na área e o fortalecimento de seu uso ao longo do tempo. A coocorrência desses três termos aponta para um possível fortalecimento do campo como aplicado e estratégico, voltado à: estudos sobre fluxos e ecossistemas informacionais digitais; criação de sistemas e políticas de gestão da informação; desenvolvimento de soluções digitais informacionais; e a competências informacionais em contextos digitais e institucionais.

Outros termos também mostraram trajetórias ascendentes, como "proposta", que cresceu de 1 para 40 ocorrências, especialmente após 2016; "arquivo", que partiu de 1

ocorrência e atingiu 28, com destaque entre 2015 e 2021; e "memória", que saltou de 1 para 26 ocorrências, com maior presença entre 2010 e 2022. Esses três termos apontam para uma possível diversificação temática e metodológica na área, que combina um maior foco propositivo (proposta), a valorização de acervos e políticas documentais (arquivo) e a atenção aos aspectos simbólicos e culturais da informação (memória).

Alguns termos apresentaram flutuações ao longo dos anos. Por exemplo, "rede" variou de 7 para 14 registros, com oscilações intermediárias; "produção" teve seu pico em 2015 (23 registros), mas caiu para 12 em 2022; e "universidade", ausente em 2004, chegou a 31 registros em 2012, oscilando nos anos seguintes entre 15 e 29 ocorrências.

Dando continuidade à análise terminológica, o Gráfico 58 apresenta a distribuição das 20 palavras mais frequentes nos títulos dos trabalhos da Ciência da Informação, segmentadas por grau acadêmico (mestrado, doutorado e mestrado profissional). Essa visualização permite observar não apenas a frequência absoluta dos termos, mas também suas associações predominantes em cada modalidade formativa.

Gráfico 58 – TOP 20 palavras mais frequentes nos trabalhos da CI por grau acadêmico; período 2004-2022

Palavras	informação	1432	582	211
	ciência	292	130	31
	estudo	520	115	120
	análise	464	138	79
	informacional	253	99	52
	proposta	88	64	101
	rede	123	52	45
	arquivo	154	55	65
	memória	155	26	52
	produção	152	60	26
	organização	147	63	25
	brasil	188	108	35
	digital	245	92	64
	sistema	173	69	56
	universidade	156	15	92
	biblioteca	360	74	193
	gestão	293	95	143
	federal	183	24	102
	conhecimento	256	145	47
	científico	252	130	46
	Mestrado		Doutorado	Mestrado Profissional
	Grau Acadêmico			

Fonte: Dados de pesquisa (2025).

Observa-se que no mestrado acadêmico, as palavras mais recorrentes são “informação”, com 1.432 ocorrências (47,57% dos títulos), seguida por “estudo” (520 ocorrências; 17,27%), “análise” (464; 15,41%), “biblioteca” (360; 11,96%) e “gestão” (293; 9,73%). Esse padrão sugere uma ênfase em abordagens empíricas e institucionais voltadas à organização e acesso à informação em ambientes bibliotecários.

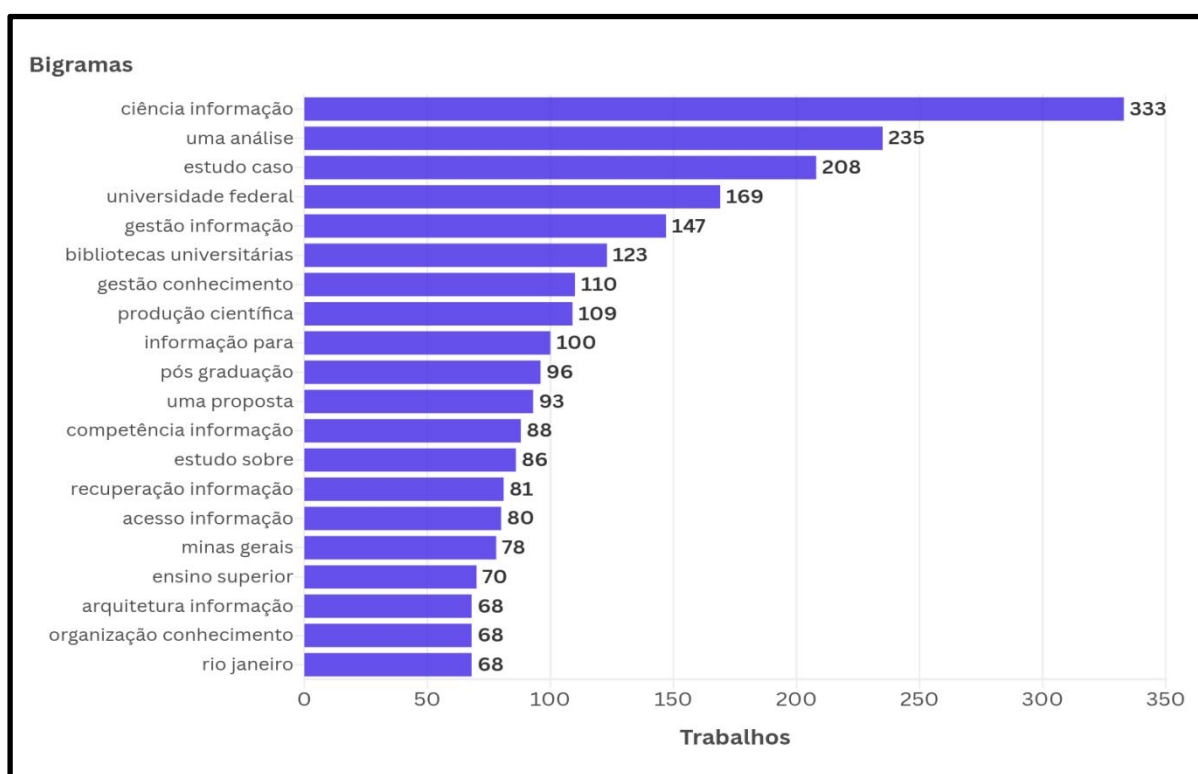
No doutorado acadêmico, os termos mais frequentes também incluem “informação”, com 582 registros (51,69%), seguido por “conhecimento” (145; 12,88%), “análise” (138; 12,26%), “ciência” (130; 11,55%) e “científico” (também 130; 11,55%). O destaque para esses termos indica maior densidade teórico-conceitual, com foco em processos de produção e validação do conhecimento científico e informacional.

Já no mestrado profissional, observa-se uma configuração distinta, voltada a aplicações práticas. Os termos mais frequentes são “informação” (211 registros; 26,48%),

“biblioteca” (193; 24,22%), “gestão” (143; 17,94%), “estudo” (120; 15,06%) e “federal” (102; 12,80%). A presença de “biblioteca” e “federal” sugere forte vínculo com instituições públicas e práticas profissionais voltadas à organização e serviços informacionais nesses espaços.

Para aprofundar a análise lexical, o Gráfico 59 apresenta os 20 bigramas²⁰ mais frequentes nos títulos dos trabalhos da área da Ciência da Informação, permitindo identificar os contextos mais recorrentes em que os termos são utilizados. A identificação dessas combinações permite observar os contextos conceituais e metodológicos mais recorrentes na produção acadêmica da área, no período analisado.

Gráfico 59 – TOP 20 bigramas mais frequentes nos títulos dos trabalhos da CI; período 2004-2022



Fonte: Dados de pesquisa (2025).

Observa-se que o bigrama mais frequente é "ciência informação", com 333 ocorrências (0,70% do total de bigramas), confirmando a centralidade dessa expressão como representação da própria área de conhecimento. A centralidade desse bigrama " nos títulos

²⁰ Um bigrama é uma sequência de duas unidades adjacentes, como letras, sílabas ou palavras, extraídas de um texto. No processamento de linguagem natural, bigramas são usados para analisar a coocorrência de palavras, ajudando a identificar padrões e relações entre termos. Essa técnica é fundamental na construção de modelos de linguagem, pois captura a estrutura e o contexto das frases, melhorando a análise textual (Silva; Souza, 2014).

pode, por um lado, reforçar a identidade da área, mas também revelar alguns possíveis problemas, tal como um possível sintoma de autorreferencialidade excessiva. Sugere que em vez de estudar fenômenos informacionais no mundo, parte da produção estuda a própria área, seus cursos, suas revistas, seus pesquisadores.

Em seguida, "uma análise" aparece em 235 títulos (0,49%), evidenciando a predominância de abordagens analíticas nas dissertações e teses. O termo "estudo caso", com 208 ocorrências (0,44%), aponta para a relevância dessa metodologia qualitativa nos trabalhos desenvolvidos.

A expressão "universidade federal", com 169 registros (0,35%), o que pode estar associado à distribuição institucional da pós-graduação da área. Já "gestão informação" (147 ocorrências, 0,31%) e "gestão conhecimento" (110 ocorrências, 0,23%) refletem duas temáticas consolidadas no campo, voltadas à organização e ao uso estratégico da informação.

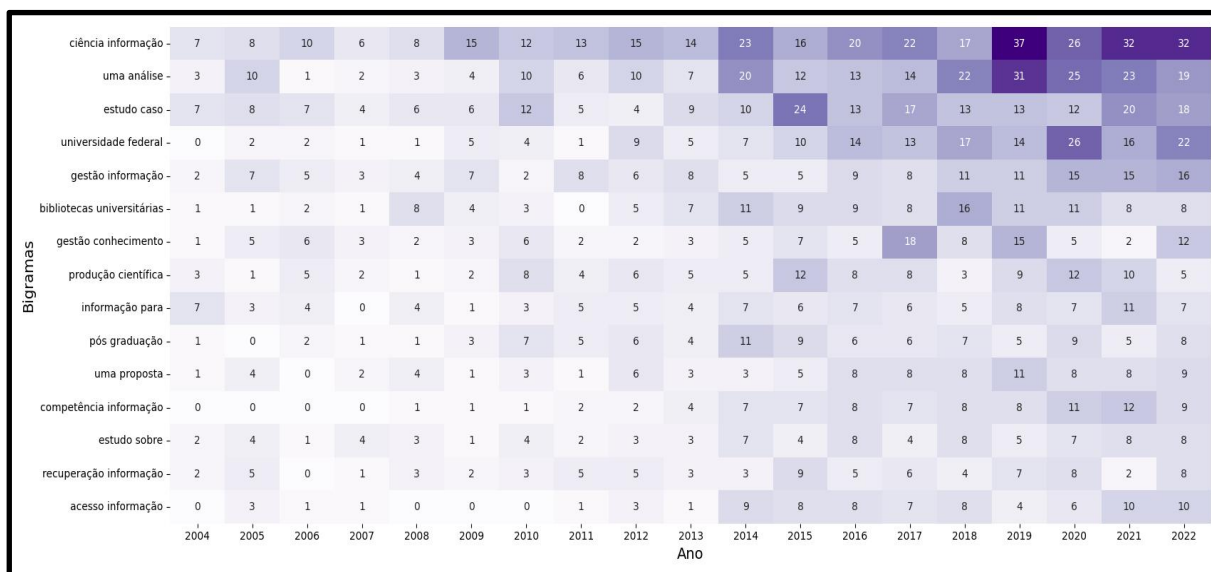
Outros bigramas em destaque incluem "bibliotecas universitárias" (123 ocorrências, 0,26%), o que sugere um volume expressivo de estudos focados nesses espaços informacionais; "produção científica" (109 ocorrências, 0,23%), voltado à análise da atividade acadêmica e da comunicação científica; e "informação para" (100 ocorrências, 0,21%), que pode remeter à função da informação em contextos como educação, saúde ou gestão pública.

Além disso, "pós-graduação" aparece em 96 títulos (0,20%), indicando o interesse em estudar a formação e produção no próprio ambiente acadêmico. Termos como "uma proposta" (93, 0,20%), "competência informação" (88, 0,18%), "estudo sobre" (86, 0,18%), "recuperação informação" (81, 0,17%) e acesso informação" (80, 0,17%), revelam enfoques em formulações teóricas e propositivas, com o desenvolvimento de habilidades informacionais e organização do conhecimento.

Outros bigramas relevantes incluem ""minas gerais" (78, 0,16%), sugerindo uma concentração de estudos regionais, "ensino superior" (70, 0,15%), "arquitetura informação" (68, 0,15%), "organização conhecimento" (68, 0,15%) e "rio janeiro" (68, 0,15%), demonstrando tanto a abrangência temática quanto recortes institucionais e geográficos presentes na produção da Ciência da Informação.

Complementando essa análise, o Gráfico 60 apresenta um *heatmap* que mostra a evolução temporal dos 20 bigramas mais frequentes nos títulos dos trabalhos. Essa visualização permite identificar variações no uso dessas expressões ao longo do tempo, evidenciando tendências temáticas e mudanças no foco da produção acadêmica da área.

Gráfico 60 – Evolução dos 20 bigramas mais frequentes nos títulos dos trabalhos da CI por ano; período 2004-2022



Fonte: Dados de pesquisa (2025).

O bigrama "ciência informação" é o mais recorrente em todo o período, passando de 7 ocorrências em 2004 (7,78% dos bigramas) para 32 em 2022 (7,21%), com picos importantes em 2019 (37 ocorrências). Isso reforça o uso do nome da própria área como elemento central nos títulos, podendo sugerir uma tendência autorreferencial. O termo "uma análise", por sua vez, demonstra crescimento contínuo, saindo de 3 ocorrências em 2004 (3,33%) para 19 em 2022 (4,28%), o que indica um predomínio de abordagens analíticas nas dissertações e teses.

A expressão "estudo caso" também apresenta presença constante e relevante, partindo de 7 ocorrências em 2004 (7,78%) e alcançando 18 registros em 2022 (4,05%), o que sinaliza a adoção recorrente desse tipo de estratégia metodológica. Já o bigrama "universidade federal", que praticamente não aparecia nos anos iniciais, cresce a partir de 2009 e alcança 22 ocorrências em 2022 (4,95%), possivelmente refletindo a expansão dos estudos em âmbito acadêmico.

Outros bigramas em ascensão incluem "gestão informação", que cresce de 2 para 16 registros (de 2,22% para 3,60%), e "bibliotecas universitárias", de 1 para 8 registros (1,11% para 1,80%). A expressão "gestão conhecimento" também ganha espaço, partindo de 1 ocorrência em 2004 (1,11%) e chegando a 12 em 2022 (2,70%).

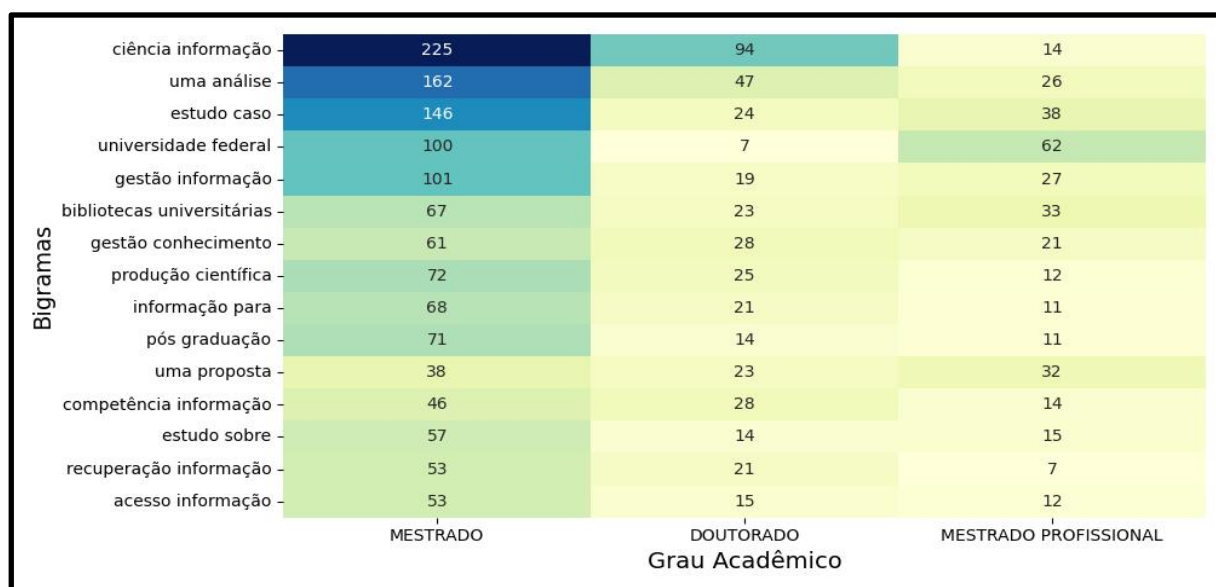
A combinação "estudo sobre" aparece 2 vezes em 2004 (2,22%) e chega a 8 ocorrências em 2022 (1,80%). O termo "recuperação informação" segue uma trajetória semelhante, aumentando de 2 para 8 ocorrências no mesmo intervalo de tempo, variando de 2,22% para 1,80%. Em seguida, "uma proposta" aumenta de 1 para 9 ocorrências

(correspondendo a 1,11% e 2,03%), enquanto "Pós-graduação" passa de 1 para 8 registros (de 1,11% para 1,80%).

Outros termos que ganharam força recentemente incluem "acesso informação", partindo de nenhuma ocorrência em 2004 para 10 registros em 2022 (2,25%), "competência informação", inexistente nos primeiros anos e com 9 ocorrências em 2022 (2,03%), e "acesso informação", que cresce de 0 para 10 registros (2,25%), refletindo debates contemporâneos sobre inclusão informacional. Esses dados evidenciam não apenas o amadurecimento da área, mas também a ampliação do escopo temático e metodológico da pesquisa em Ciência da Informação.

Dando continuidade à análise terminológica, o Gráfico 61 apresenta um *heatmap* com a distribuição dos 20 bigramas mais frequentes nos títulos dos trabalhos da Ciência da Informação, segmentados por grau acadêmico. A visualização possibilita identificar como os diferentes níveis da pós-graduação — mestrado acadêmico, doutorado acadêmico e mestrado profissional — se diferenciam em termos da terminologia predominante nos títulos das dissertações e teses, destacando padrões específicos de enfoque conceitual e temático.

Gráfico 61 – Distribuição dos 20 bigramas mais frequentes nos títulos dos trabalhos da CI por grau; período 2004-2022



Fonte: Dados de pesquisa (2025).

No mestrado acadêmico, o bigrama mais recorrente é "ciência informação", com 225 ocorrências (7,47%), seguido por "uma análise", com 162 ocorrências (5,38%) e "estudo caso", com 146 ocorrências (4,85%), revelando a centralidade da área e o predomínio de abordagens empíricas e analíticas. Bigramas como "gestão informação", com 101 registros

(3,35%) e "universidade federal", com 100 registros (3,32%) também ganham destaque, sinalizando o vínculo institucional e a ênfase em temas aplicados à gestão da informação.

No doutorado acadêmico, a liderança também é de "ciência informação", com 94 ocorrências (8,35%), reforçando o caráter autorreferencial já apontado em análises anteriores. Em seguida, aparecem "uma análise" (47, 4,17%), "gestão conhecimento" (28, 2,49%), "competência informação" (28, 2,49%) e "produção científica" (25, 2,22%), sugerindo uma maior densidade teórica e foco em temas epistemológicos, formação de competências e avaliação da ciência.

No mestrado profissional, observa-se um padrão distinto. O bigrama mais frequente é "universidade federal", com 62 ocorrências (7,78%), o que pode refletir tanto o perfil institucional das pesquisas desenvolvidas quanto a influência dos Planos de Desenvolvimento Institucional (PDI), que frequentemente estabelecem metas de qualificação para servidores e docentes, impulsionando a realização de trabalhos voltados às demandas e realidades das IES. Em seguida, destacam-se "estudo caso" (38 – 4,77%), "bibliotecas universitárias" (33 – 4,14%), "uma proposta" (32 – 4,02%) e "gestão informação" (27 – 3,39%). Esses termos apontam para uma forte orientação prática e institucional, com foco em ambientes informacionais específicos e na proposição de soluções ou melhorias de processos.

Passando a análise para os resumos dos trabalhos, o Gráfico 62 apresenta um *heatmap* que exibe a coocorrência das 20 palavras mais utilizadas nos resumos dos trabalhos da Ciência da Informação, permitindo a identificação das associações mais recorrentes entre os termos.

Gráfico 62 – Coocorrência das 20 palavras mais frequentes nos resumos dos trabalhos da CI; período 2004-2022

biblioteca	0	257	180	223	379	214	264	312	315	334	428	321	386	891	810	677	577	567	571	602
digital	257	0	207	259	287	260	234	280	353	312	427	376	418	790	801	604	534	539	501	530
gestão	180	207	0	372	217	246	377	476	351	433	673	385	551	1034	986	801	739	711	680	688
documento	223	259	372	0	216	322	361	369	267	440	583	481	482	1036	966	786	759	642	592	671
usuário	379	287	217	216	0	199	405	370	450	352	541	408	453	951	1004	769	662	660	638	630
científico	214	260	246	322	199	0	307	568	286	371	531	770	522	1127	1004	857	811	754	721	677
sistema	264	234	377	361	405	307	0	487	347	449	637	493	508	1079	1089	797	767	765	679	697
conhecimento	312	280	476	369	370	568	487	0	526	614	910	835	738	1553	1499	1226	1118	1035	905	952
informacional	315	353	351	267	450	286	347	526	0	470	761	598	615	1251	1336	1007	870	809	808	780
trabalho	334	312	433	440	352	371	449	614	470	0	825	605	664	1401	1399	1059	983	939	858	931
processo	428	427	673	583	541	531	637	910	761	825	0	884	962	1973	1934	1528	1393	1328	1189	1277
ciência	321	376	385	481	408	770	493	835	598	605	884	0	766	1740	1813	1358	1260	1095	1017	1026
meio	386	418	551	482	453	522	508	738	615	664	962	766	0	1735	1629	1336	1282	1145	1130	1141
pesquisa	891	790	1034	1036	951	1127	1079	1553	1251	1401	1973	1740	1735	0	3450	2785	2604	2442	2276	2397
informação	810	801	986	966	1004	1004	1089	1499	1336	1399	1934	1813	1629	3450	0	2649	2431	2245	2096	2148
estudo	677	604	801	786	769	857	797	1226	1007	1059	1528	1358	1336	2785	2649	0	1986	1845	1720	1800
análise	577	534	739	759	662	811	767	1118	870	983	1393	1260	1282	2604	2431	1986	0	1712	1742	1668
resultado	567	539	711	642	660	754	765	1035	809	939	1328	1095	1145	2442	2245	1845	1712	0	1602	1624
dado	571	501	680	592	638	721	679	905	808	858	1189	1017	1130	2276	2096	1720	1742	1602	0	1513
objetivo	602	530	688	671	630	677	697	952	780	931	1277	1026	1141	2397	2148	1800	1668	1624	1513	0
biblioteca																				
digital																				
gestão																				
documento																				
usuário																				
científico																				
sistema																				
conhecimento																				
informacional																				
trabalho																				
processo																				
ciência																				
meio																				
pesquisa																				
informação																				
estudo																				
análise																				
resultado																				
dado																				
objetivo																				

Fonte: Dados de pesquisa (2025).

A combinação mais frequente é entre os termos "pesquisa" e "informação", com 3.450 coocorrências (69,92%), refletindo a centralidade da informação como objeto de investigação na área. Em seguida, destacam-se os pares "pesquisa" e "estudo" (2.785 ocorrências, 56,45%) e "informação" e "estudo" (2.649 ocorrências, 53,69%), reforçando o foco em processos investigativos sobre a informação.

Outras associações relevantes incluem "pesquisa" e "análise" (2.604 ocorrências, 52,78%) e "pesquisa" e "resultado" (2.442 ocorrências, 49,49%), evidenciando a presença de métodos analíticos e resultados empíricos nos resumos. A relação entre "informação" e "análise" (2.431 ocorrências, 49,27%) e "pesquisa" e "objetivo" (2.397 ocorrências, 48,58%) também se destacam, indicando uma estrutura comum de descrição das investigações.

Termos como "dado", "objetivo", "processo" e "ciência" aparecem frequentemente associados tanto a "pesquisa" quanto a "informação", formando um campo lexical que remete à prática científica e aos métodos de produção e organização do conhecimento. Por exemplo,

"pesquisa" e "dado" coocorrem em 2.276 resumos (46,13%), "informação" e "resultado" em 2.245 (45,50%) e "informação" e "objetivo" em 2.148 (43,53%).

Outras associações expressivas incluem "estudo" e "análise" (1.986, 40,25%), "processo" e "pesquisa" (1.973, 39,99%), "processo" e "informação" (1.934, 39,20%), "estudo" e "resultado" (1.845, 37,39%), "ciência" e "informação" (1.813, 36,75%), "estudo" e "objetivo" (1.800, 36,48%), "análise" e "dado" (1.742, 35,31%) e "ciência" e "pesquisa" (1.740, 35,27%), revelando o protagonismo desses termos na composição dos resumos. A presença de "meio" em diversas combinações (ex.: “meio” e “pesquisa”, 1.735 ocorrências, ou 35,16%) aponta para um possível foco nas mediações e ambientes nos quais a informação circula ou é estruturada.

Dando continuidade à análise lexical dos resumos, o Gráfico 63 apresenta um *heatmap* com a distribuição anual das 20 palavras mais frequentes nos resumos dos trabalhos. Essa visualização permite identificar tendências temáticas ao longo do tempo, apontando para a emergência, consolidação ou declínio de determinados termos, o que pode refletir transformações nos interesses de pesquisa, nas abordagens metodológicas e nas prioridades da área em diferentes contextos históricos e institucionais.

Gráfico 63 – TOP 20 palavras mais frequentes nos resumos dos trabalhos da CI por ano; período 2004-2022

Palavras	Ano																			
	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	
biblioteca -	62	53	72	59	127	102	139	114	217	230	278	367	320	288	453	408	341	431	419	
digital -	40	72	42	62	64	61	169	105	170	100	102	160	175	218	270	268	311	300	344	
gestão -	39	87	91	71	58	72	131	117	79	136	137	236	227	346	273	381	284	334	410	
documento -	29	45	54	56	41	74	60	86	128	192	164	227	198	261	250	378	271	358	335	
usuário -	39	63	61	71	118	67	109	111	159	200	164	141	185	175	211	200	173	259	179	
científico -	49	54	97	52	86	149	129	132	204	140	185	287	219	232	256	366	386	378	415	
sistema -	69	121	71	59	112	91	85	113	111	137	143	201	219	211	208	270	216	252	234	
conhecimento -	99	141	142	101	107	164	199	123	177	188	256	280	263	394	330	409	299	346	384	
informacional -	53	100	59	71	112	121	159	192	177	209	197	186	175	186	263	336	270	344	310	
trabalho -	77	90	81	78	81	70	90	94	115	103	165	180	151	184	193	218	205	191	224	
processo -	81	139	162	111	135	158	178	181	220	252	324	309	264	388	342	474	416	422	471	
ciência -	43	108	83	114	127	161	154	192	206	193	234	255	255	294	267	450	431	437	387	
meio -	21	51	49	35	78	54	93	92	103	116	163	222	174	228	264	314	260	282	325	
pesquisa -	122	172	213	209	276	292	363	409	491	510	663	798	769	947	971	1211	1044	1228	1323	
informação -	391	666	606	522	603	668	710	939	917	977	1161	1258	1197	1223	1321	1728	1332	1830	1497	
estudo -	89	146	132	140	195	179	212	220	263	286	382	418	352	452	541	560	479	550	619	
análise -	72	110	124	113	157	156	195	208	245	270	388	395	348	406	482	576	446	576	582	
resultado -	51	73	76	81	77	92	98	129	187	183	219	279	252	289	318	335	319	388	370	
dado -	43	50	83	68	98	100	134	146	204	195	284	366	314	425	447	639	588	777	743	
objetivo -	35	46	45	52	68	74	97	95	119	127	151	183	200	230	269	301	298	348	362	

Fonte: Dados de pesquisa (2025).

Nos anos mais recentes, observa-se o crescimento expressivo de termos. O termo "gestão", por exemplo, passou de 284 ocorrências em 2020 para 410 em 2022, representando um aumento de 2,55% em relação ao total de trabalhos. A palavra "conhecimento" também apresentou crescimento, passando de 299 para 384 registros, com uma variação de 1,72%. O termo "processo", que já figurava entre os mais frequentes, registrou 416 ocorrências em 2020 e 471 em 2022, o que equivale a 1,11% do total. A palavra "meio" também ganhou espaço, crescendo de 260 em 2020 para 325 registros em 2022, representando 1,32% do conjunto analisado.

O termo "gestão" aumentou de 284 ocorrências em 2020 para 410 em 2022, correspondendo a um crescimento de 2,55% em relação ao total de trabalhos. A palavra "conhecimento" também apresentou crescimento no período recente, passando de 299 registros em 2020 para 384 em 2022, representando um aumento de 1,72%. A expressão "processo" registrou 416 ocorrências em 2020 e aumentou para 471 em 2022, correspondendo

a 1,11% do total. O termo "meio" também apresentou crescimento, passando de 260 registros em 2020 para 325 em 2022, o que representa um aumento de 1,32% em relação ao total de trabalhos analisados.

Outros termos com tendência de alta incluem "pesquisa", que passou de 1.044 em 2020 para 1.323 ocorrências em 2022 (crescimento de 5,65%), e "estudo", que passou de 479 para 619 registros no mesmo intervalo (aumento de 2,83%). O termo "análise" evoluiu de 446 para 582 ocorrências, o que representa um crescimento de 2,75%. Já a palavra "objetivo" foi de 298 para 362 registros, crescendo 1,30%.

Esses aumentos sugerem uma maior ênfase em processos investigativos e analíticos, além de uma intensificação da linguagem voltada à gestão do conhecimento, à produção científica e à sistematização metodológica nos resumos dos trabalhos da área.

Continuando a análise dos resumos, o Gráfico 64 apresenta um *heatmap* que exhibe a distribuição das 20 palavras mais frequentes, segmentadas por grau acadêmico. A análise permite identificar diferenças na terminologia utilizada nos resumos de dissertações e teses, destacando padrões específicos para cada nível da pós-graduação.

Gráfico 64 – TOP 20 palavras mais frequentes nos resumos dos trabalhos da CI por grau acadêmico; período 2004-2022

Palavras	Mestrado	Doutorado	Mestrado Profissional
	2553	682	1245
	1824	756	453
	1868	711	930
	1852	757	598
	1714	616	355
	2198	1231	387
	1667	757	499
	2420	1453	529
	2074	958	488
	1575	540	475
	2916	1299	812
	2760	1392	239
	1721	759	444
	6984	3060	1967
	12360	5118	2068
	3836	1512	867
	3509	1551	789
	2234	1007	575
	3251	1511	942
	1835	708	557
	Mestrado	Doutorado	Mestrado Profissional

Fonte: Dados de pesquisa (2025).

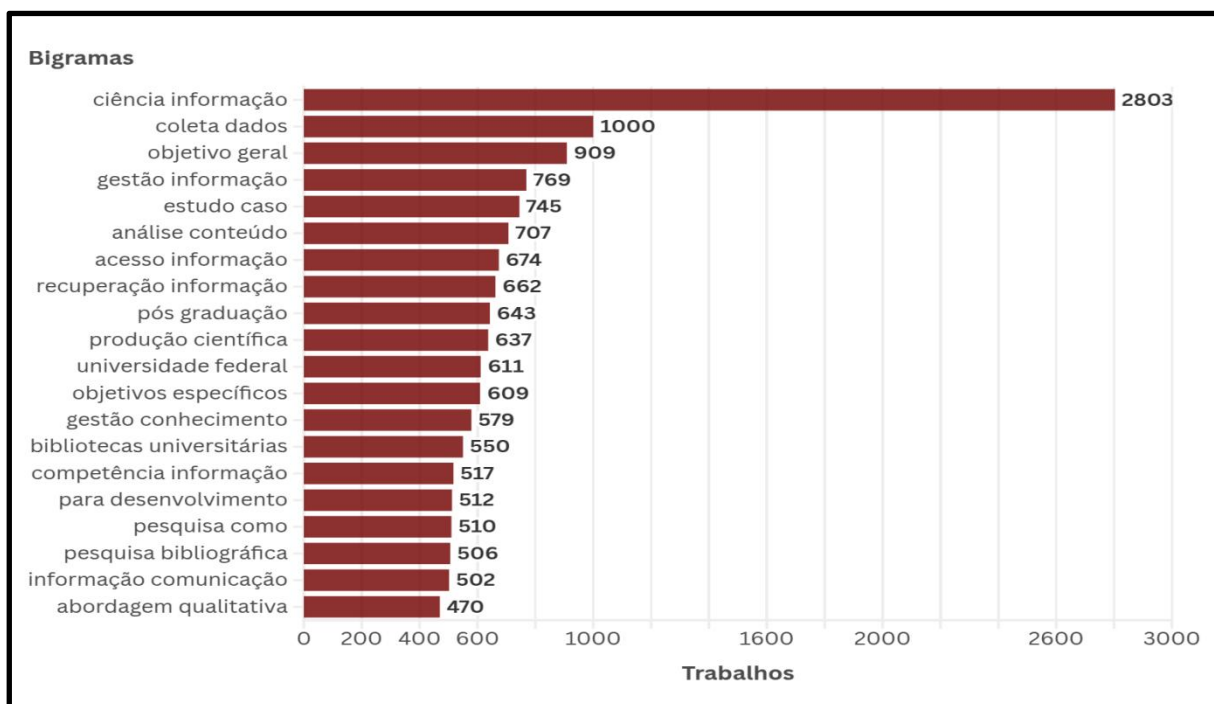
No mestrado acadêmico, a palavra "informação" aparece com maior destaque, totalizando 12.360 ocorrências (25,05%). Em seguida, aparecem "pesquisa", com 6.984 ocorrências (14,15%), "estudo", com 3.836 registros (7,78%), "análise", com 3.509 ocorrências (7,11%), e "dado", com 3.251 menções (6,59%). Esses termos reforçam a ênfase em atividades investigativas e metodologias aplicadas que caracterizam este grau.

No doutorado acadêmico, com menor proporção de trabalhos, "informação" também se destaca com 5.118 ocorrências (10,37%), seguida de "pesquisa" (3.060 ocorrências; 6,20%), "análise" (1.551, 3,14%), "estudo" (1.512, 3,06%) e "dado" (1.511, 3,06%). Essa modalidade também mantém o foco temático característico da área.

No mestrado profissional, a palavra mais frequente é também "informação", com 2.068 registros (4,19%), seguida de "pesquisa", com 1.967 ocorrências (3,99%). "Biblioteca" aparece em 1.245 resumos (2,52%), seguida por "dado" (942 ocorrências; 1,91%) e "gestão" (930 registros; 1,89%), que ganham destaque, revelando uma maior aproximação com demandas práticas e institucionais nessa modalidade.

Prosseguindo ainda com os resumos, o gráfico 65 apresenta os 20 bigramas mais frequentes nos trabalhos, permitindo identificar as associações mais recorrentes entre os termos utilizados. A distribuição dos bigramas revela padrões linguísticos que ajudam a compreender os enfoques temáticos predominantes na produção acadêmica.

Gráfico 65 – TOP 20 bigramas mais frequentes nos resumos dos trabalhos da CI; período 2004-2022



Fonte: Dados de pesquisa (2025).

O bigrama mais recorrente é "ciência informação", com 2.803 ocorrências, o que representa 56,82% do total de trabalhos analisados. Em seguida, destaca-se "coleta dados", com 1.000 registros (20,26%), sinalizando o foco em procedimentos metodológicos de coleta empírica. A expressão "objetivo geral" aparece em 909 resumos (18,42%), frequentemente relacionada à estrutura padrão de apresentação de projetos e dissertações. O termo "gestão informação" ocorre em 769 resumos (15,59%), indicando estudos que abordam um campo teórico-prático voltado ao tratamento sistemático da informação. Já "estudo caso", em 745 trabalhos (15,10%) e "análise conteúdo" (707 ocorrências; 14,33%), apontam para a popularidade desses tipos de abordagens metodológicas.

Outros pares de palavras com alta frequência incluem "acesso informação" (674 registros; 13,66%) e "recuperação informação" (662 ocorrências; 13,42%), refletindo o interesse da área por temas ligados ao tratamento e uso da informação. A expressão "pós-

graduação" aparece em 643 trabalhos (13,03%), enquanto "produção científica" totaliza 637 ocorrências (12,91%), o que pode indicar reflexões sobre a própria atividade de pesquisa.

Também são frequentes bigramas como "universidade federal" (611 registros; 12,38%), "objetivos específicos" (609 ocorrências; 12,34%) e "gestão conhecimento" (579 registros; 11,73%). As expressões "bibliotecas universitárias" (550; 11,15%) e "competência informação" (517; 10,47%) evidenciam o diálogo da área com o campo bibliotecário e educacional. Outros bigramas relevantes incluem "para desenvolvimento" (512; 10,38%), "pesquisa como" (510; 10,34%), "pesquisa bibliográfica" (506; 10,26%), "informação comunicação" (502; 10,17%) e "abordagem qualitativa" (470 ocorrências; 9,52%).

Ainda no exame dos resumos, o Gráfico 66 exibe os 20 bigramas mais frequentes ao longo do período analisado, possibilitando uma análise da evolução terminológica nos trabalhos da área de Ciência da Informação.

Gráfico 66 – TOP 20 bigramas mais frequentes nos resumos dos trabalhos da CI por ano; período 2004-2022

Bigramas	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
ciência informação -	33	84	58	87	76	98	104	143	127	134	157	153	153	201	153	295	237	276	234
coleta dados -	11	4	20	12	23	25	38	39	40	44	50	72	67	75	83	96	88	106	107
objetivo geral -	3	6	5	7	6	11	18	15	26	29	43	44	46	58	86	118	109	125	154
gestão informação -	15	27	17	12	19	33	31	33	25	33	29	37	71	41	42	65	75	79	85
estudo caso -	12	11	18	10	21	14	29	22	25	40	42	60	38	64	57	63	65	75	79
análise conteúdo -	1	5	10	3	16	14	19	18	20	28	42	37	47	46	65	81	87	79	89
acesso informação -	0	14	8	8	13	7	10	11	16	25	48	57	52	45	66	78	50	96	70
recuperação informação -	8	23	15	23	18	32	22	34	29	37	22	58	51	43	35	58	52	54	48
pós graduação -	5	8	8	10	16	25	37	24	45	22	53	42	36	38	59	56	44	44	71
produção científica -	5	2	17	9	7	20	35	15	39	15	18	62	38	43	35	67	78	67	65
universidade federal -	0	6	5	5	8	19	10	14	35	18	35	39	40	50	70	78	56	54	69
objetivos específicos -	4	4	4	4	7	13	16	9	16	11	25	30	35	33	71	84	66	84	93
gestão conhecimento -	9	24	32	13	13	7	46	7	9	12	27	48	23	93	58	53	30	10	65
bibliotecas universitárias -	4	6	10	7	24	9	25	2	30	47	38	44	46	30	53	49	37	48	41
competência informação -	0	1	0	0	6	8	12	11	11	29	45	36	54	31	22	69	45	78	59

Fonte: Dados de pesquisa (2025).

O bigrama "ciência informação" mantém-se como o mais recorrente ao longo do tempo, saltando de 33 ocorrências em 2004 para 234 em 2022, o que representa um crescimento de 609,09%. A expressão "coleta dados" apresentou um aumento ainda mais expressivo, de 11 para 107 registros, marcando um crescimento de 872,73%, refletindo o fortalecimento de abordagens empíricas nos trabalhos da área. Já o termo "objetivo geral" apresentou um crescimento notável de 5.033,33%, passando de 3 ocorrências em 2004 para

154 em 2022, e "objetivos específicos", que passou de 4 para 93 ocorrências (2.225%), provavelmente relacionados à formalização da estrutura textual dos resumos.

Outras associações também evidenciam tendências importantes: "gestão informação" passou de 15 para 85 ocorrências (466,67%), e "estudo caso" de 12 para 79 (558,33%), indicando o crescimento do interesse por abordagens voltadas à prática e à aplicação. A expressão "análise conteúdo" teve um crescimento expressivo de 8.800,00%, partindo de 1 ocorrência em 2004 para 89 em 2022, apontando para a consolidação dessa técnica metodológica nas pesquisas da área.

Alguns bigramas que não apareciam em 2004 também ganharam destaque ao longo do tempo, como "acesso informação", que alcançou 70 registros em 2022, "universidade federal", com 69 registros em 2022, e "competência informação", que não ocorria em 2004 e passou para 59 registros em 2022.

Por sua vez, o termo "recuperação informação" cresceu de 8 para 48 ocorrências (500%), "pós-graduação" de 5 para 71 (1.320%), e "produção científica" de 5 para 65 (1.200%). O termo "gestão conhecimento" evoluiu de 9 para 65 registros (622,22%), e a expressão "bibliotecas universitárias" também apresentou crescimento, passando de 4 para 41 registros (925%). Por fim, o termo "competência informação", ausente em 2004, alcançou 59 ocorrências em 2022, indicando sua ascensão como conceito relevante nas discussões da área.

Prosseguindo com a análise dos resumos, o Gráfico 67 apresenta os 20 bigramas mais, agora segmentados por grau acadêmico (mestrado acadêmico, doutorado acadêmico e mestrado profissional), permitindo a análise da distribuição terminológica em cada nível da pós-graduação.

Gráfico 67 – TOP 20 bigramas mais frequentes nos resumos dos trabalhos da CI por grau acadêmico; período 2004-2022

Bigramas	ciência informação	1856	829	118
	coleta dados	594	206	200
	objetivo geral	491	224	194
	gestão informação	462	133	174
	estudo caso	463	139	143
	análise conteúdo	423	173	111
	acesso informação	426	147	101
	recuperação informação	400	197	65
	pós graduação	420	134	89
	produção científica	372	200	65
	universidade federal	384	63	164
	objetivos específicos	357	119	133
	gestão conhecimento	327	150	102
	bibliotecas universitárias	305	125	120
	competência informação	280	185	52
		MESTRADO	DOCTORADO	MESTRADO PROFISSIONAL
		Grau Acadêmico		

Fonte: Dados de pesquisa (2025).

No mestrado, o bigrama "ciência informação" aparece como o mais recorrente, com 1.856 ocorrências, correspondendo a aproximadamente 61,64% dos trabalhos desse nível. Em seguida, destacam-se os bigramas "coleta dados" (594 ocorrências, 19,73%), "objetivo geral" (491 ocorrências, 16,31%), "estudo caso" (463 ocorrências, 15,38%) e "gestão informação" (462 ocorrências, 15,34%), sugerindo uma forte presença de pesquisas empíricas com ênfase em coleta de dados.

No doutorado, embora os temas se mantenham, observa-se um aprofundamento em aspectos investigativos e conceituais. O bigrama "ciência informação" permanece na liderança, com 829 ocorrências (73,62%), seguido de "objetivo geral" (224 ocorrências, 19,89%) e "coleta dados" (206 ocorrências, 18,29%). Destacam-se ainda "produção científica" (200 ocorrências, 17,76%) e "recuperação informação" (197 ocorrências, 17,50%), apontando para a valorização de temas relacionados à geração, organização e acesso ao conhecimento científico.

Já no mestrado profissional, nota-se um padrão distinto, com maior ênfase na aplicação prática dos conceitos. O termo mais frequente é "coleta dados", com 200 registros (25,09%), seguido de "objetivo geral" (194 ocorrências, 24,34%) e "gestão informação" (174 ocorrências, 21,83%). A expressão "universidade federal" aparece com destaque (164 ocorrências, 20,58%), assim como "estudo caso" (143 ocorrências, 17,94%), o que sugere forte ligação entre os projetos desenvolvidos e os contextos institucionais nos quais os

estudantes estão inseridos — possivelmente em sintonia com os PDIs, que incentivam a qualificação técnico-científica voltada à resolução de problemas concretos.

Com base nas análises apresentadas ao longo da seção 5.7.5, foi possível identificar características dos trabalhos defendidos entre 2004 e 2022, com padrões terminológicos e institucionais recorrentes nos títulos e resumos dos trabalhos defendidos na área da Ciência da Informação. Essas informações fornecem subsídios importantes para a compreensão da identidade científica da área e das dinâmicas de sua produção pós-graduada no período analisado.

Na sequência, a seção 5.7.6 apresenta os resultados da caracterização dos programas, dentro do recorte dos dados em Ciência da Informação, oferecendo um aprofundamento específico das informações relacionadas aos Programas de Pós-Graduação *stricto sensu* vinculados à área no sistema da CAPES.

5.7.6 Resultados da etapa do Recorte dos dados da Ciência da Informação no DF4_programas_CI

O conjunto de dados DF4_programas_CI difere do apresentado na seção 5.7.2, referente ao DF4_programas_BR, por abranger exclusivamente os programas da área de Ciência da Informação que permaneceram em funcionamento no sistema da CAPES entre os anos de 2004 e 2022. Os quatro programas desativados durante esse período não estão contemplados neste recorte, o que garante um foco analítico sobre os programas ativos ao longo da série histórica considerada (2004-2022), que foi possível com o cruzamento de dados.

Essa filtragem permite uma abordagem mais específica, voltada para a caracterização dos programas em funcionamento, com base em variáveis como natureza (*status*) jurídica (considerando: federal, estadual, municipal e privada), dependência administrativa (considerando: pública e privada), município de funcionamento, modalidade (graus acadêmico e profissional) e ano de início das ofertas de mestrado e doutorado. Além disso, o conjunto possibilita a análise dos conceitos atribuídos pela CAPES aos programas da área, com recorte temporal entre os anos de 2013 e 2022.

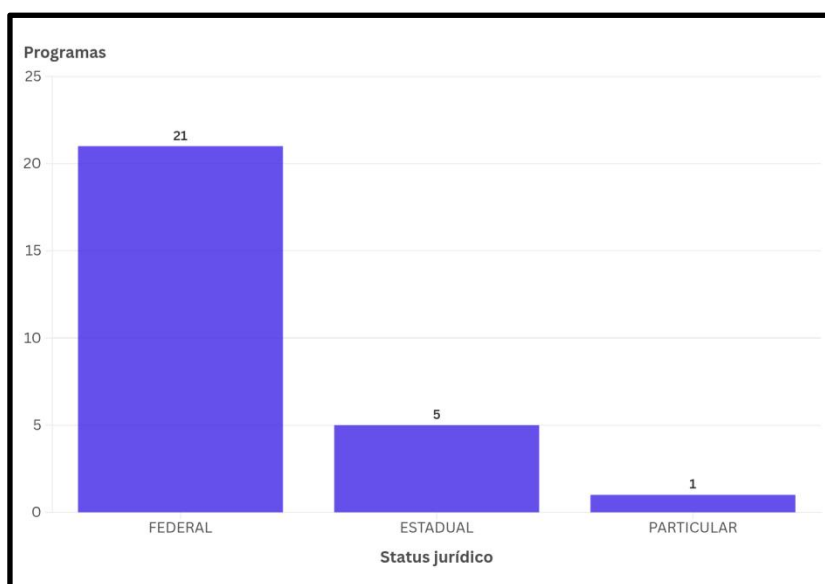
Cabe esclarecer que, embora o escopo do estudo contemple dados da pós-graduação brasileira entre 2004 e 2022, os dados estruturados sobre todos os programas de pós-graduação disponíveis nas bases acessadas se restringem ao período de 2013 a 2022. Devido ao volume expressivo dessas informações no âmbito nacional, não foi viável desenvolver uma

análise detalhada sobre o conjunto DF4_programas_BR (seção 5.7.2) no escopo desta pesquisa.

A escolha metodológica por concentrar a análise na área de Ciência da Informação justifica-se, portanto, não apenas pela viabilidade do tratamento dos dados, mas também pela possibilidade de exploração mais aprofundada de suas características. Embora os dados da Ciência da Informação também estejam organizados majoritariamente entre 2013 e 2022, identificou-se, nesse conjunto, informações adicionais sobre os programas, como o ano de início das ofertas de mestrado e doutorado, mesmo quando anteriores a 2013. Dessa maneira, a delimitação ao DF4_programas_CI viabilizou uma investigação aprofundada, alinhada tanto aos propósitos analíticos do estudo quanto às restrições operacionais identificadas no tratamento dos dados.

Iniciando a análise dos programas em Ciência da Informação, o Gráfico 68 apresenta a distribuição dos programas em funcionamento na área no período 2013-2022, classificados segundo sua natureza (*status*) jurídica. Embora o recorte temporal da análise apresentada compreenda esse período, não se espera que a natureza jurídica dos programas tenha sofrido alterações ao longo do tempo analisado (2004-2022), uma vez que esse atributo tende a se manter estável após a criação e credenciamento dos cursos.

Gráfico 68 – Distribuição dos programas em funcionamento da CI por natureza (status) jurídica; dados do período 2013-2022



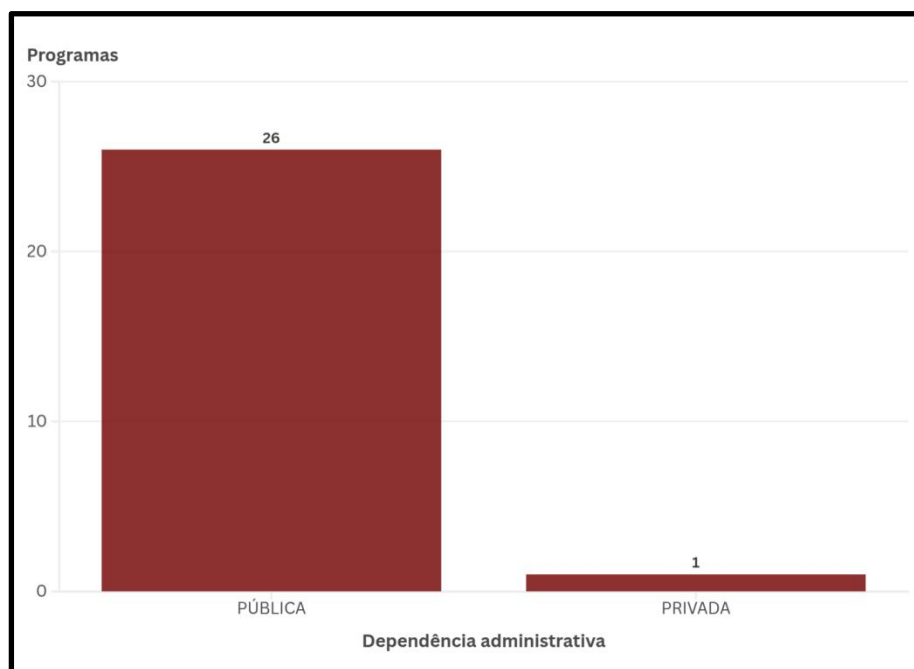
Fonte: Dados de pesquisa (2025).

Observa-se que a maior parte dos programas está vinculada a instituições federais, totalizando 21 registros, o que representa 77,78% do total. Em seguida, aparecem as instituições estaduais, com cinco programas ativos (18,52%). Já as instituições particulares

possuem apenas um programa em funcionamento no período analisado, correspondendo a 3,70% do total de programas da área.

Prosseguindo a análise, o Gráfico 69 apresenta a distribuição dos programas em funcionamento na área da Ciência da Informação, classificados por dependência administrativa. Assim como a natureza jurídica, esse elemento não costuma sofrer alterações ao longo do tempo, pois está diretamente vinculado à estrutura institucional da universidade que oferta o programa. Portanto, ainda que os dados se refiram ao período de 2013 a 2022, presume-se a estabilidade dessa característica desde o credenciamento dos cursos.

Gráfico 69 – Distribuição dos programas em funcionamento da CI por dependência administrativa; dados do período 2013-2022



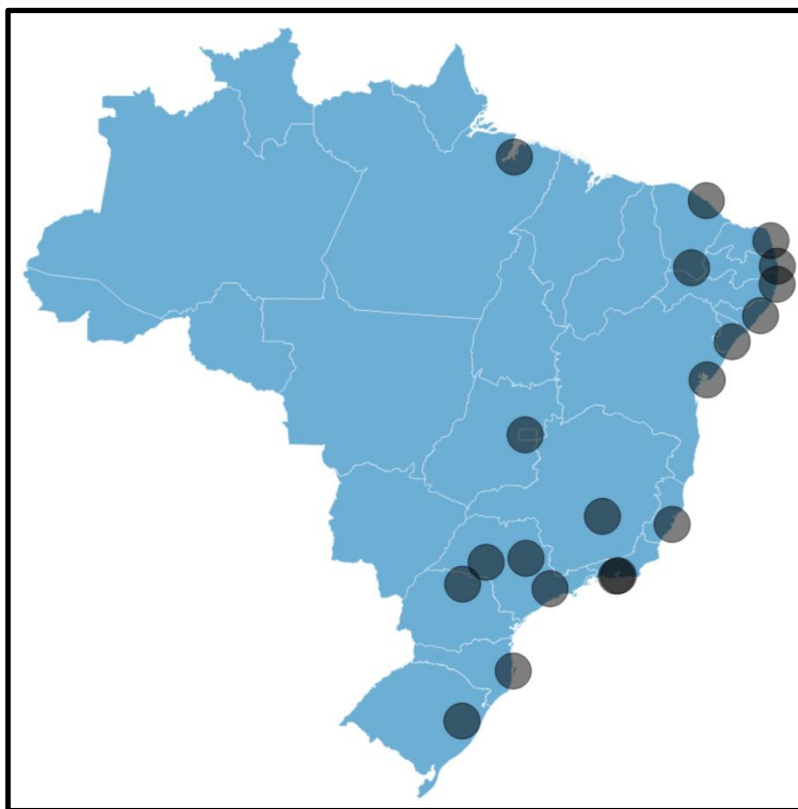
Fonte: Dados de pesquisa (2025).

Observa-se que a maioria dos programas em Ciência da Informação está vinculada a instituições públicas, totalizando 26 registros, o que corresponde a 96,30% do total. Apenas um programa pertence a uma instituição privada (FUMEC), representando 3,70% do conjunto analisado.

Assim como a natureza jurídica e a dependência administrativa, esse atributo também tende a permanecer estável ao longo do tempo, uma vez que está diretamente associado à infraestrutura física e à localização institucional da universidade credenciada. Portanto, ainda que os dados se refiram ao período de 2013 a 2022, presume-se que a distribuição geográfica dos programas tenha se mantido constante durante esse intervalo.

O Gráfico 70 apresenta a distribuição geográfica dos programas da Ciência da Informação no Brasil.

Gráfico 70 – Localização das cidades dos programas em funcionamento da CI; dados do período 2013-2022



Fonte: Dados de pesquisa (2025).

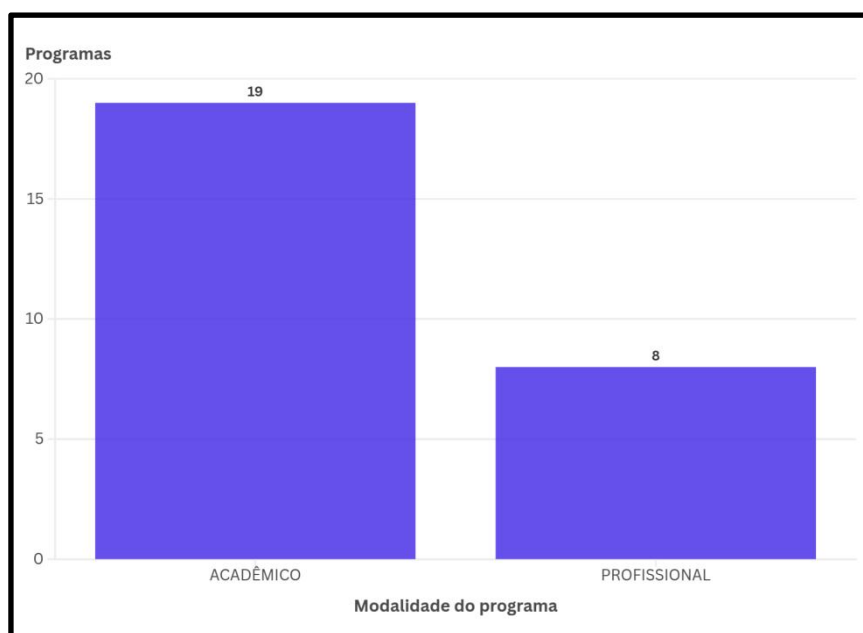
Observa-se que a maioria dos programas em funcionamento está situada em capitais estaduais, com destaque para o Rio de Janeiro, que concentra quatro programas, seguido por Belo Horizonte, com três. São Paulo e Florianópolis abrigam dois programas cada. Outras capitais como Fortaleza, Vitória, Maceió, Belém, Brasília, Natal, João Pessoa, Salvador, Recife e Porto Alegre contam com um programa cada. Esse padrão evidencia a concentração da oferta em grandes centros urbanos, tradicionalmente associados a instituições de ensino superior mais consolidadas.

No entanto, nota-se também a presença de programas em cidades não capitais, o que sinaliza um esforço de descentralização da pós-graduação *stricto sensu* na área. São Cristóvão, São Carlos, Juazeiro do Norte, Londrina, Marília e Niterói são exemplos de municípios que, embora não sejam capitais, sediam programas ativos em Ciência da Informação, contribuindo para a interiorização e diversificação da oferta no território nacional. Essas localidades compartilham algumas características em comum, como a presença de universidades públicas

consolidadas (a exemplo da UFS, UFSCar, UFCA, UEL, UNESP e UFF), assim como são cidades de médios portes, que possuem um papel regional relevante ao concentrarem atividades educacionais, científicas e culturais.

Prosseguindo com a análise, o Gráfico 71 apresenta a distribuição dos programas em funcionamento, organizadas modalidade: acadêmico e profissional. Ainda que a modalidade de um programa possa, em tese, ser alterada ao longo do tempo, esse tipo de mudança é pode ser considerada rara. A definição da modalidade está vinculada ao projeto político-pedagógico original aprovado pela CAPES e à vocação institucional da universidade promotora. Por isso, presume-se que essa característica tenha se mantido estável.

Gráfico 71 – Distribuição dos programas em funcionamento da CI por modalidade; dados do período 2013-2022



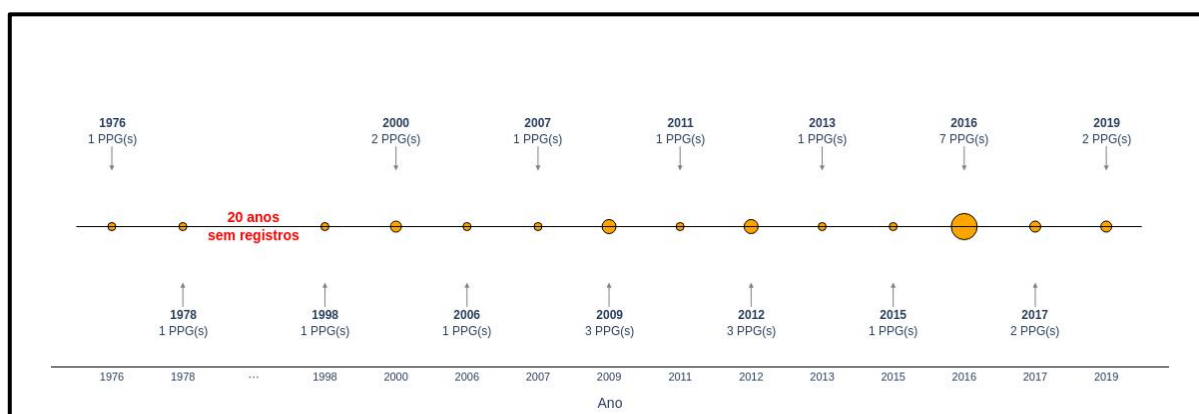
Fonte: Dados de pesquisa (2025).

Observa-se que a maior parte dos programas pertence à modalidade acadêmica, com 19 registros, o que corresponde a 70% do total analisado. Já os programas profissionais somam 8 registros, representando 30% dos programas em funcionamento entre 2013 e 2022.

A predominância da modalidade acadêmica está relacionada à consolidação histórica dos programas de pós-graduação voltados para a formação de pesquisadores e docentes. Por outro lado, a presença crescente de programas profissionais — instituídos com maior intensidade a partir da década de 2010 — reflete políticas de expansão da pós-graduação orientadas à aplicação do conhecimento científico em contextos práticos e ao atendimento de demandas específicas do mercado e da gestão pública.

Dando continuidade à análise, o Gráfico 73 apresenta a distribuição dos programas em funcionamento na área da Ciência da Informação, segundo o ano de início, permitindo observar a dinâmica de criação dos cursos ao longo do tempo. Como dito na introdução desta seção, nos dados de 2013-2022 constavam as datas de início dos programas de Ciências da Informação em funcionamento.

Gráfico 72 – Distribuição do ano início dos programas em funcionamento da CI; dados do período 2013-2022



Fonte: Dados de pesquisa (2025).

A trajetória da área de Ciência da Informação revela uma expansão progressiva, marcada por momentos distintos de crescimento. Os programas pioneiros foram criados nos anos de 1976, 1978 e 1998, cada um representando 3,70% do total de programas ativos.

A década de 2000 trouxe um novo impulso, com dois programas iniciados em 2000 (7,41%) e um em 2006 (3,70%). Entre 2007 e 2012, houve acréscimos graduais, com destaque para os anos de 2009 e 2012, quando três novos programas foram criados em cada ano, totalizando 11,11% do conjunto atual em cada um desses períodos, sendo que no ano de 2011 houve um único programa (3,70%).

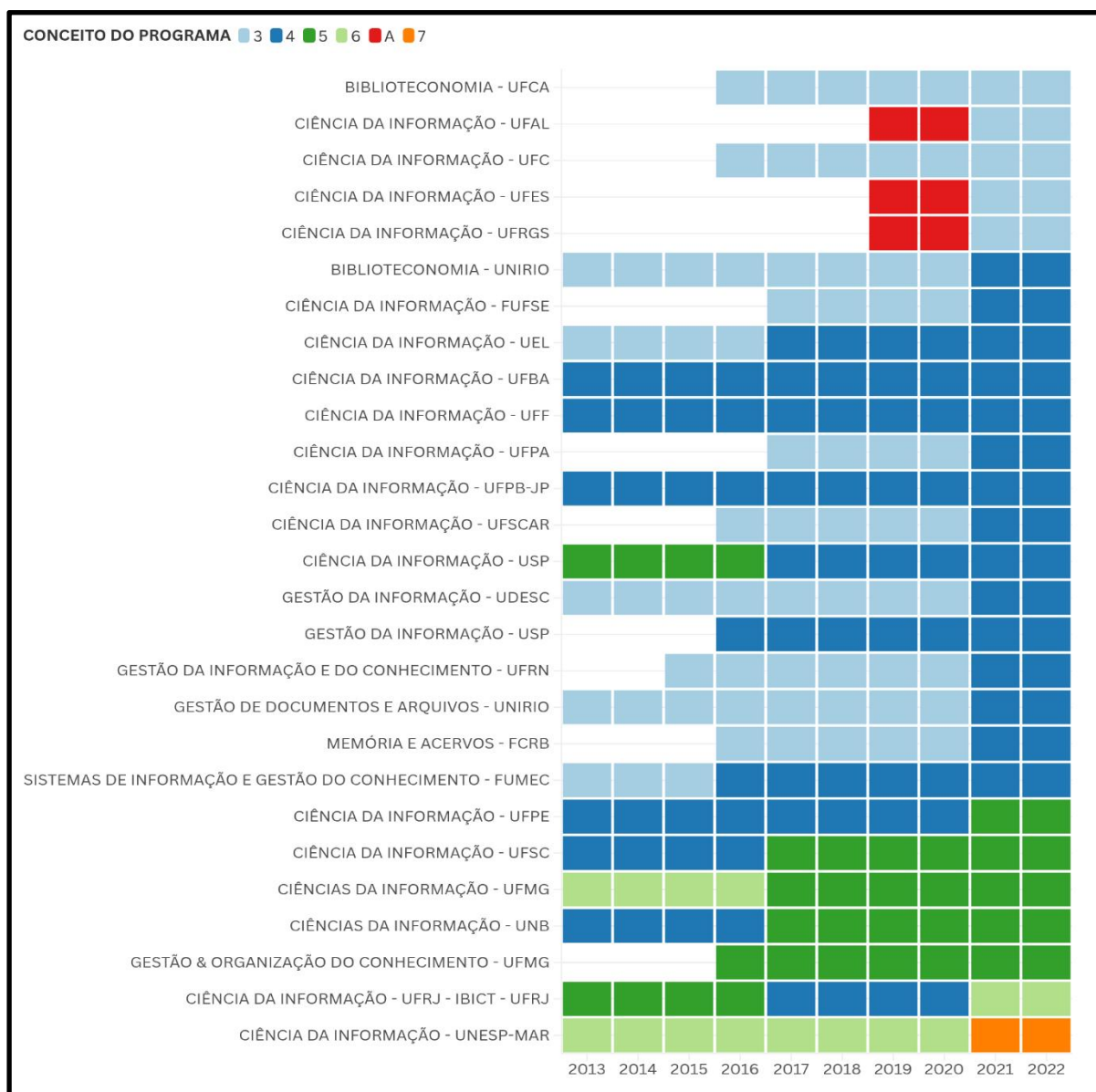
Depois, em 2013 e 2015, houve a criação de um programa (3,70%) para cada. E, em 2016, ocorreu a maior expansão em 2016, com sete novos programas foram criados, o que representa 25,93% do total de programas ativos — a maior concentração registrada em um único ano. Esse crescimento expressivo reflete, possivelmente, políticas públicas voltadas à interiorização da pós-graduação, à diversificação temática e à criação de programas profissionais.

Nos anos seguintes, observa-se uma desaceleração no ritmo de criação em relação a 2016: dois em 2017 (7,41%) e dois em 2019 (7,41%). Ainda assim, esses anos mantêm a continuidade da expansão da área, consolidando sua presença em diferentes regiões do país.

Essa distribuição revela três grandes momentos de crescimento: a fundação dos programas pioneiros na década de 1970; uma retomada entre o final dos anos 1990 e início dos 2000; e, por fim, um salto expressivo entre 2009 e 2016, provavelmente impulsionado por políticas públicas voltadas à interiorização e diversificação da pós-graduação, bem como à criação de programas profissionais.

Continuando a análise, o Gráfico 73 apresenta a distribuição dos conceitos atribuídos aos programas de pós-graduação da área de Ciência da Informação pela CAPES, no período de 2013 a 2022. Ressalta-se que após 2012, a CAPES adotou o sistema de avaliações quadrienais para os programas de pós-graduação no Brasil. No período de 2013 a 2022, os quadriênios de avaliação foram dois: 2013-2016 e 2017-2020. Contudo, tem-se que levar em consideração que há avaliações parciais anuais da Capes, com atribuição de notas.

Gráfico 73 – Distribuição das notas dos programas de Ciência da Informação; dados do período 2013-2022



Fonte: Dados de pesquisa (2025).

A série histórica revela que o programa da UNESP (Marília) se destacou pela consistência, com conceito 6 de 2013 a 2020. Em 2021, o programa foi o único a alcançar o conceito 7, avaliação máxima concedida pela CAPES, destacando-se nacionalmente pela excelência de sua produção científica, formação discente e impacto acadêmico.

Destaque também para o programa Ciência da Informação da UFRJ-IBICT, que iniciou o período com conceito 5 (2013–2016), sofreu uma redução para nota 4 entre 2017 e 2020, mas apresentou recuperação expressiva em 2021, atingindo conceito 6, sinalizando avanço qualitativo nas avaliações mais recentes.

O conceito 5 aparece em programas que alcançaram um estágio avançado de maturidade. Entre os programas esse conceito em 2022, destaca-se o programa Ciência da Informação da UFMG, que obteve conceito 6 de 2013 a 2016, porém, foi reavaliado com conceito 5 a partir de 2017 mantendo essa nota até 2022. Por sua vez, o programa da UnB ficou com conceito 4 de 2013 a 2016, e ascendeu ao conceito 5 em 2017, mantendo essa avaliação até 2022. Por sua vez, o programa da UFPE repetiu o mesmo desempenho do programa da UnB. Distinção pode ser dada ao programa da UFMG – Gestão e Organização do Conhecimento, que obteve o conceito 5 desde a sua implementação, em 2016, mantendo essa nota até 2022. Em contrapartida, o programa Ciência da Informação da USP ficou com conceito 5 entre 2013 a 2016, e sofreu redução na nota, estando com conceito 4 de 2017 a 2022.

A maior parte dos programas mais novos inicia sua trajetória com conceito 3, como é o caso dos programas da FCRB, FUFSE, UDESC, UEL, UFC, UFCA, UFPA, UFRN, UFSCAR e os dois do UNIRIO. Com o tempo, observa-se a progressão de alguns deles para o conceito 4, como os da FCRB, FUFSE, UDESC, UEL, UFPA, UFRN, UFPE, UFSCAR e os dois do UNIRIO. Essa progressão indica consolidação da produção científica, da formação discente e da infraestrutura dos programas. Outros, já iniciaram o programa com o conceito 4 e permanecem nele até 2022, como é o caso dos programas da FUMEC, UFBA, UFF, UFPB e o programa Gestão da Informação da USP.

No outro extremo, alguns programas permaneceram autorizados com o conceito “A”. Esse conceito aparece principalmente nos documentos de reconhecimento de novos cursos pela CAPES, que pode indicar programas recém-criados ou em fase de consolidação, antes do conceito numérico a ser atribuído pela avaliação quadrienal. Nesse caso estão os programas da UFRGS, UFES e UFAL, ambos com conceito “A” entre 2019 e 2020, alterando a nota para 3 entre 2021 e 2022.

Antes de finalizar a análise dos programas de pós-graduação em Ciência da Informação, decidiu-se cruzar os dados disponíveis no DF3_trabalhos_CI (seção 5.7.5), que tem informações sobre a produção de teses e dissertações no período de 2004 a 2022 (todo o período analisado neste estudo), com os dados dos programas (DF4_programas_CI). Isso foi realizado para que seja possível a consolidação da caracterização dos programas juntamente com a suas respectivas produções no período 2004-2022. Desse modo, a Tabela 37 apresenta dados consolidados sobre os programas de pós-graduação em Ciência da Informação, considerando a modalidade, o nome do programa, a IES de vinculação, o número total de

trabalhos defendidos (2004-2022), o ano de início do programa, a idade do programa e a média de produção por ano de existência.

Tabela 37 – Características e produção dos programas de pós-graduação em Ciência da Informação no Brasil; dados do período 2004-2022

Índice	Modalidade	Nome programa	IES	Trabalhos	Início	Idade	Trabalhos por ano
1	PROFISSIONAL	MEMÓRIA E ACERVOS	FCRB	55	2016	6	9.17
2	PROFISSIONAL	CIÊNCIA DA INFORMAÇÃO	FUFSE	56	2017	5	11.20
3	ACADÊMICO	CIÊNCIA DA INFORMAÇÃO	UNB	513	1978	44	11.66
4	ACADÊMICO	CIÊNCIA DA INFORMAÇÃO	USP	267	2006	16	16.69
5	PROFISSIONAL	GESTÃO DA INFORMAÇÃO	USP	25	2016	6	4.17
6	PROFISSIONAL	GESTÃO DA INFORMAÇÃO	UFSC	106	2013	9	11.78
7	ACADÊMICO	CIÊNCIA DA INFORMAÇÃO	UEL	104	2012	10	10.40
8	ACADÊMICO	CIÊNCIA DA INFORMAÇÃO	UNESP	490	1998	24	20.42
9	ACADÊMICO	CIÊNCIA DA INFORMAÇÃO	UFBA	250	2000	22	11.36
10	ACADÊMICO	CIÊNCIA DA INFORMAÇÃO	UFPB	333	2007	15	22.20
11	ACADÊMICO	CIÊNCIA DA INFORMAÇÃO	UFAL	22	2016	6	3.67
12	ACADÊMICO	CIÊNCIA DA INFORMAÇÃO	UFMG	540	1976	46	11.74
13	ACADÊMICO	GESTÃO & ORGANIZAÇÃO DO CONHECIMENTO	UFMG	143	2016	6	23.83
14	ACADÊMICO	CIÊNCIA DA INFORMAÇÃO	UFPE	173	2009	13	13.31
15	ACADÊMICO	CIÊNCIA DA INFORMAÇÃO	UFSC	300	2000	22	13.64
16	ACADÊMICO	CIÊNCIA DA INFORMAÇÃO	UFSCAR	50	2016	6	8.33
17	PROFISSIONAL	BIBLIOTECONOMIA	UFCA	79	2016	6	13.17
18	ACADÊMICO	CIÊNCIA DA INFORMAÇÃO	UFC	43	2016	6	7.17
19	ACADÊMICO	CIÊNCIA DA INFORMAÇÃO	UFES	17	2019	3	5.67
20	PROFISSIONAL	BIBLIOTECONOMIA	UFRJ	164	2012	10	16.40
21	PROFISSIONAL	GESTÃO DE DOCUMENTOS E ARQUIVOS	UNIRIO	85	2012	10	8.50
22	ACADÊMICO	CIÊNCIA DA INFORMAÇÃO	UFPA	53	2017	5	10.60
23	ACADÊMICO	CIÊNCIA DA INFORMAÇÃO - UFRJ – IBICT	UFRJ	296	2009	13	22.77
24	PROFISSIONAL	CIÊNCIA DA	UFRN	69	2015	7	9.86

Índice	Modalidade	Nome programa	IES	Trabalhos	Início	Idade	Trabalhos por ano
		INFORMAÇÃO					
25	ACADÊMICO	CIÊNCIA DA INFORMAÇÃO	UFRGS	36	2019	3	12.00
26	ACADÊMICO	CIÊNCIA DA INFORMAÇÃO	UFF	178	2009	13	13.69
		TECNOLOGIA DA INFORMAÇÃO E					
27	ACADÊMICO	COMUNICAÇÃO E GESTÃO DO CONHECIMENTO	FUMEC	267	2011	11	24.27

Fonte: Dados de pesquisa (2025).

Observa-se que o programa Ciência da Informação da UFMG é o mais antigo entre os analisados, tendo iniciado suas atividades em 1976. Com 46 anos de funcionamento, acumula 540 trabalhos defendidos e uma média anual de 11,74 trabalhos.

O segundo programa mais antigo é o da UnB, criado em 1978, com 513 trabalhos defendidos e média de 11,66 trabalhos por ano. A Unesp, por sua vez, criou seu programa em 1998, apresentando um desempenho expressivo com 490 trabalhos defendidos e uma média anual de 20,42. Outros programas relevantes iniciados no início dos anos 2000 incluem os da UFBA e UFSC (ambos em 2000), com 250 e 300 trabalhos, e médias de 11,36 e 13,64, respectivamente.

Entre os programas criados a partir de 2009, destaca-se o da UFRJ-IBICT²¹, com 296 trabalhos defendidos e uma alta média de 22,77 por ano, desde sua criação. A UFPE, iniciada também em 2009, tem 173 trabalhos e uma média de 13,31. No âmbito da modalidade profissional, o pioneiro é o programa de Biblioteconomia da UFRJ, criado em 2012, com 164 trabalhos defendidos e média de 16,40 por ano.

Programas mais recentes, como o Memória e Acervos da FCRB (iniciado em 2016) e o Ciência da Informação da FUFSE (iniciado em 2017), apresentam médias anuais de 9,17 e 11,20, respectivamente. O programa Gestão & Organização do Conhecimento da UFMG, de modalidade acadêmica, criado em 2016, destaca-se por sua produtividade, com média de 23,83 trabalhos por ano. Já o programa Tecnologia da Informação e Comunicação e Gestão do Conhecimento da FUMEC, criado em 2011, possui a maior média anual do conjunto: 24,27 trabalhos.

²¹ Os dados apresentados referem-se ao período analisado do programa de pós-graduação UFRJ-IBICT atualmente em atividade. Ressalta-se que existiu um programa anterior vinculado ao IBICT em associação com a UFF, o qual já se encontra extinto. Em virtude de sua descontinuidade, este programa não foi considerado na presente análise.

A comparação entre os programas revela variações expressivas nas médias de produção anual, independentemente do tempo de existência ou da modalidade (acadêmica ou profissional). Enquanto os programas mais antigos tendem a apresentar estabilidade produtiva, os programas profissionais — em geral mais recentes — mostram médias elevadas e, em alguns casos, níveis de produção comparáveis ou superiores aos dos programas tradicionais. Isso indica não apenas o dinamismo dos novos programas, mas também a consolidação da área em diferentes configurações institucionais.

E, por fim, ainda com base no cruzamento entre os dados do DF3_trabalhos_CI (seção 5.7.5) e do DF4_programas_CI, a Tabela 38 apresenta o ano de início dos cursos de mestrado e doutorado para os programas de pós-graduação da área de Ciência da Informação, assim como o volume de teses e dissertações defendidas em cada curso no período de 2004 a 2022.

Tabela 38 – Panorama dos programas de Ciência da Informação com início do mestrado e doutorado e produção acadêmica; dados do período 2004-2022

Índice	Nome Programa	IES	Trabalhos	Modalidade do programa	Início Mestrado	Início Doutorado
1	MEMÓRIA E ACERVOS	FUNDACAO CASA DE RUI BARBOSA	55	PROFISSIONAL	2016	-
2	CIÊNCIA DA INFORMAÇÃO	FUNDAÇÃO UNIVERSIDADE FEDERAL DE SERGIPE	56	PROFISSIONAL	2017	-
3	CIÊNCIA DA INFORMAÇÃO	UNIVERSIDADE DE BRASÍLIA	513	ACADÊMICO	1978	1992
4	CIÊNCIA DA INFORMAÇÃO	UNIVERSIDADE DE SÃO PAULO	267	ACADÊMICO	2006	2006
5	GESTÃO DA INFORMAÇÃO	UNIVERSIDADE DE SÃO PAULO	25	PROFISSIONAL	2016	-
6	GESTÃO DA INFORMAÇÃO	UNIVERSIDADE DO ESTADO DE SANTA CATARINA	106	PROFISSIONAL	2013	-
7	CIÊNCIA DA INFORMAÇÃO	UNIVERSIDADE ESTADUAL DE LONDRINA	104	ACADÊMICO	2012	2018
8	CIÊNCIA DA INFORMAÇÃO	UNIVERSIDADE ESTADUAL PAULISTA JÚLIO DE MESQUITA FILHO	490	ACADÊMICO	1998	2005
9	CIÊNCIA DA INFORMAÇÃO	UNIVERSIDADE FEDERAL DA BAHIA	250	ACADÊMICO	2000	2011
10	CIÊNCIA DA INFORMAÇÃO	UNIVERSIDADE FEDERAL DA PARAIBA	333	ACADÊMICO	2007	2012
11	CIÊNCIA DA INFORMAÇÃO	UNIVERSIDADE FEDERAL DE ALAGOAS	22	ACADÊMICO	2019	-

Índice	Nome Programa	IES	Trabalhos	Modalidade do programa	Início Mestrado	Início Doutorado
12	CIÊNCIA DA INFORMAÇÃO	UNIVERSIDADE FEDERAL DE MINAS GERAIS	540	ACADÊMICO	1976	1997
13	GESTÃO & ORGANIZAÇÃO DO CONHECIMENTO	UNIVERSIDADE FEDERAL DE MINAS GERAIS	143	ACADÊMICO	2016	2016
14	CIÊNCIA DA INFORMAÇÃO	UNIVERSIDADE FEDERAL DE PERNAMBUCO	173	ACADÊMICO	2009	-
15	CIÊNCIA DA INFORMAÇÃO	UNIVERSIDADE FEDERAL DE SANTA CATARINA	300	ACADÊMICO	2000	2013
16	CIÊNCIA DA INFORMAÇÃO	UNIVERSIDADE FEDERAL DE SÃO CARLOS	50	ACADÊMICO	2016	-
17	BIBLIOTECONOMIA	UNIVERSIDADE FEDERAL DO CARIRI	79	PROFISSIONAL	2016	-
18	CIÊNCIA DA INFORMAÇÃO	UNIVERSIDADE FEDERAL DO CEARÁ	43	ACADÊMICO	2016	-
19	CIÊNCIA DA INFORMAÇÃO	UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO	17	ACADÊMICO	2019	-
20	BIBLIOTECONOMIA	UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO	164	PROFISSIONAL	2012	-
21	GESTÃO DE DOCUMENTOS E ARQUIVOS	UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO	85	PROFISSIONAL	2012	-
22	CIÊNCIA DA INFORMAÇÃO	UNIVERSIDADE FEDERAL DO PARÁ	53	ACADÊMICO	2017	-
23	CIÊNCIA DA INFORMAÇÃO - UFRJ - IBICT	UNIVERSIDADE FEDERAL DO RIO DE JANEIRO	296	ACADÊMICO	2009	2009
24	CIÊNCIA DA INFORMAÇÃO	UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE	69	PROFISSIONAL	2015	-
25	CIÊNCIA DA INFORMAÇÃO	UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL	36	ACADÊMICO	2019	-
26	CIÊNCIA DA INFORMAÇÃO	UNIVERSIDADE FEDERAL FLUMINENSE	178	ACADÊMICO	2009	2014
27	TECNOLOGIA DA INFORMAÇÃO E COMUNICAÇÃO E GESTÃO DO CONHECIMENTO	UNIVERSIDADE FUMEC	267	ACADÊMICO	2011	2016

Fonte: Dados de pesquisa (2025).

Observa-se que algumas instituições estruturaram seus programas desde o início para ofertar os dois níveis de formação. É o caso do programa de Ciência da Informação da USP, implantado em 2006; do programa Ciência da Informação – UFRJ-IBICT, criado em 2009; e do programa Gestão & Organização do Conhecimento da UFMG, iniciado em 2016 — todos com ingresso simultâneo de turmas de mestrado e doutorado.

Em contrapartida, outros programas adotaram um modelo de implementação progressiva. A UnB, por exemplo, introduziu o doutorado 14 anos após a criação do mestrado, originalmente implantado em 1978. Já o programa da UEL apresentou uma das transições mais ágeis, incorporando o doutorado apenas seis anos após o início do mestrado. Casos intermediários incluem o programa da UFBA, cujo doutorado foi estabelecido 11 anos após o mestrado.

Ainda há programas em funcionamento que não oferecem doutorado. É o caso do programa de Biblioteconomia da UNIRIO, criado em 2012, que, até 2022, não havia expandido para o nível doutoral. Esse cenário revela a diversidade de estratégias institucionais na construção da pós-graduação *stricto sensu* na área, refletindo tanto políticas institucionais quanto dinâmicas regionais e temáticas.

Finalizada a análise dos dados dos programas, a próxima seção apresenta os resultados do recorte dos dados da Ciência da Informação no DF5_docentes_CI, com foco no perfil e na atuação do corpo docente dos programas.

5.7.7 Resultados da etapa do Recorte dos dados da Ciência da Informação no DF5_docentes_CI

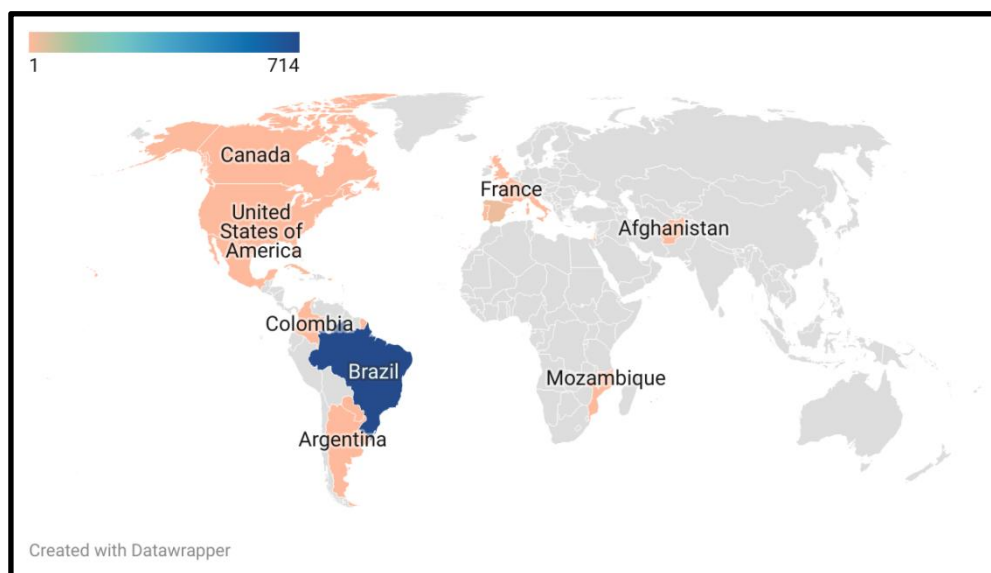
O conjunto de dados do DF5_docentes_CI, recortado especificamente para a área da Ciência da Informação, contempla informações sobre docentes vinculados a programas de pós-graduação *stricto sensu* no Brasil entre os anos de 2004 e 2022. A base é composta por 6.057 registros, que correspondem a 766 docentes distintos, permitindo uma análise robusta da trajetória acadêmica e profissional dos professores que atuaram na área ao longo do período.

As variáveis analisadas incluem a nacionalidade, ano de nascimento, categoria institucional na IES, nome e país da última titulação, ano da titulação, nome da instituição formadora, vínculo atual, área básica da formação e grande área do conhecimento. Esses campos oferecem subsídios para análises tanto individuais quanto agregadas, possibilitando o

mapeamento da qualificação docente, da mobilidade acadêmica e da composição institucional da área da Ciência da Informação.

Inicia-se a análise com o Gráfico 74, que apresenta a distribuição da nacionalidade dos docentes da Ciência da Informação no período analisado.

Gráfico 74 – Mapa geral da nacionalidade dos docentes da CI; período 2004-2022

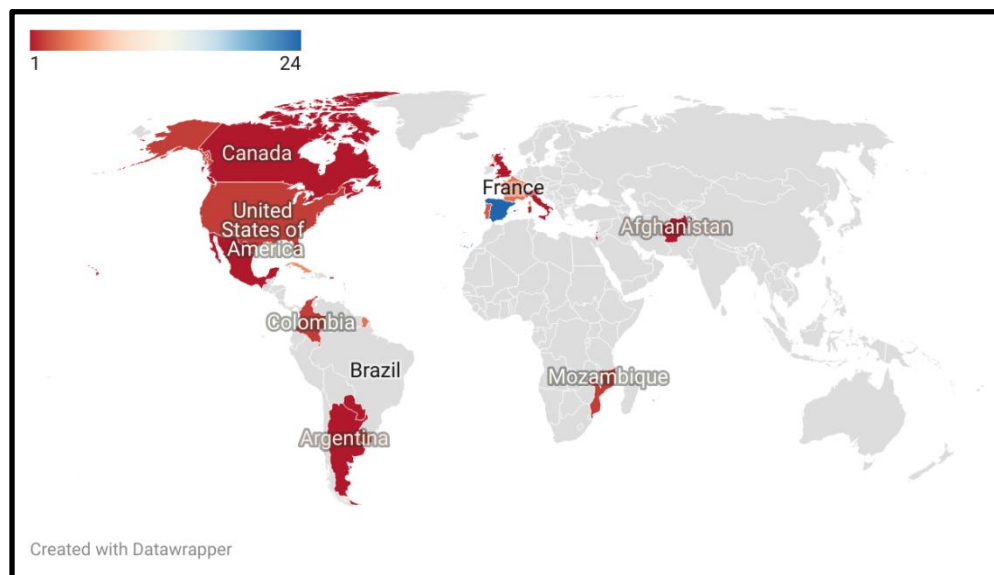


Fonte: Dados de pesquisa (2025).

A análise da nacionalidade dos docentes revela uma forte predominância de docentes brasileiros, que representam 93,21% do total (714 indivíduos). Os docentes estrangeiros correspondem a 6,79% da amostra (52 registros), evidenciando uma participação minoritária, porém plural em termos de origem geográfica.

O Gráfico 75 complementa esta análise ao apresentar exclusivamente a distribuição dos docentes estrangeiros na área da Ciência da Informação, excluindo os registros de nacionalidade brasileira.

Gráfico 75 – Mapa da nacionalidade dos docentes estrangeiros da CI (exceto Brasil); período 2004-2022



Fonte: Dados de pesquisa (2025).

A partir dessa visualização, observa-se que a maior concentração de docentes estrangeiros provém da Espanha, com 24 registros (3,13%). Em seguida, destacam-se Cuba, com cinco docentes (0,65%), e França, com quatro registros (0,52%). Portugal aparece com três docentes (0,39%), enquanto Moçambique, Estados Unidos, Colômbia e Uruguai contabilizam dois registros cada (0,26%).

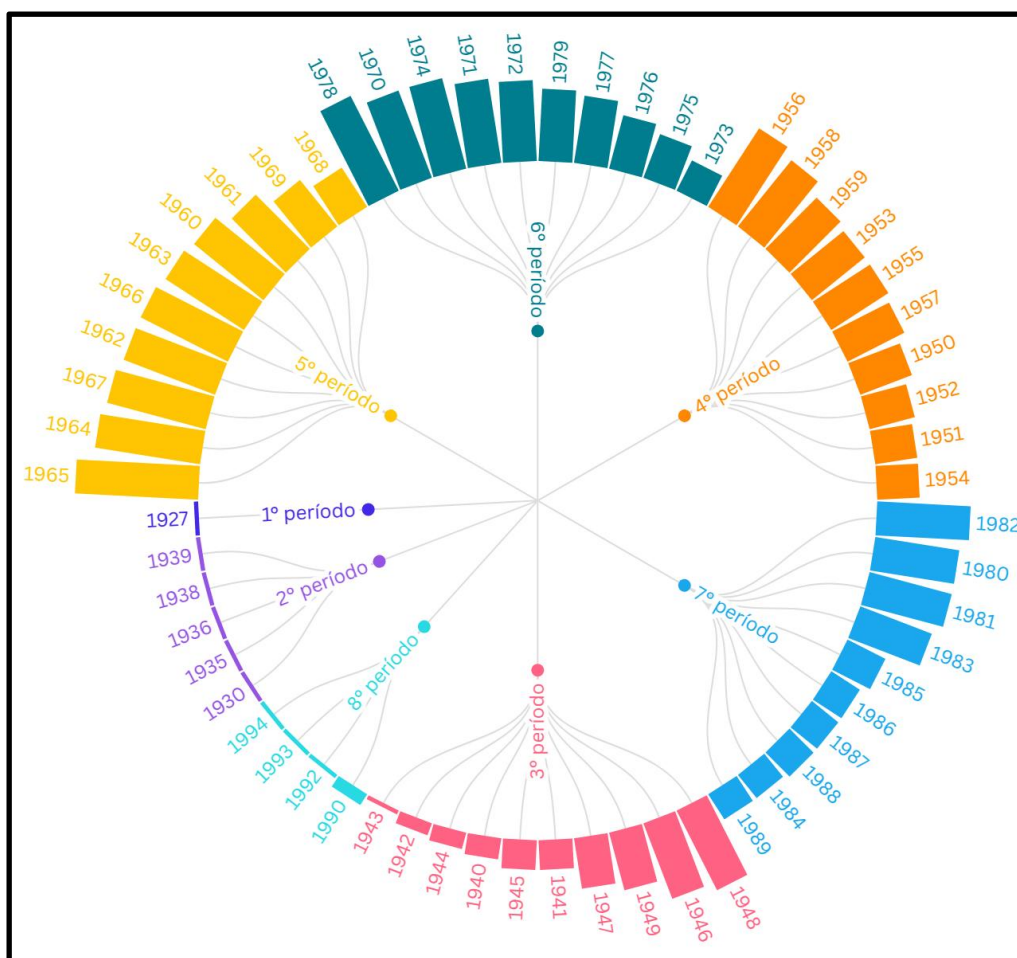
Entre os países de origem dos docentes estrangeiros, destaca-se a Espanha, com 24 registros (3,13%), seguida por Cuba com cinco docentes (0,65%) e França, com quatro registros (0,52%). Portugal aparece com três docentes (0,39%).

Outras nacionalidades presentes, embora com apenas um registro cada (0,13%), incluem Afeganistão, Argentina, Canadá, Inglaterra, Israel, Itália, México e Paraguai. Essa distribuição evidencia que, embora a inserção de docentes estrangeiros seja limitada em termos quantitativos, ela reflete uma diversidade de vínculos internacionais, concentrada especialmente em países ibero-americanos e lusófonos, sugerindo afinidades linguísticas e acadêmicas que favorecem cooperação e mobilidade.

Dando continuidade à análise do perfil dos docentes, o Gráfico 76 apresenta a distribuição dos anos de nascimento dos professores vinculados aos programas de pós-graduação em Ciência da Informação no Brasil. A visualização foi estruturada em formato radial e agrupada por oito faixas, a saber: 1º período: só 1927, 2º período: 1930 a 1939, 3º período: 1940 a 1949, 4º período: 1950 a 1959, 5º período: 1960 a 1969, 6º período: 1970 a 1979, 7º período: 1980 a 1989 e 8º período: 1990 a 1994. Com isso, é possível identificar a

predominância geracional entre os docentes da área e observar possíveis tendências de renovação ou envelhecimento do corpo docente ao longo do tempo.

Gráfico 76 – Distribuição dos docentes da Ciência da Informação por faixas de nascimento; período 2004-2022



Fonte: Dados de pesquisa (2025).

O 1º período, que compreende apenas o ano de 1927, registra 1 docente. O 2º período (1930–1939) contabiliza 4 registros (0,52%). Já o 3º período (1940–1949) reúne 82 docentes, representando 10,69% do total. Os anos com maior concentração nesse grupo são 1948 (21 docentes, 2,74%) e 1946 (19 docentes, 2,48%).

O 4º período (1950–1959) soma 140 registros (18,26%), com destaque para 1956 (22 docentes, 2,87%) e 1958 (20 docentes, 2,61%). O 5º período (1960–1969) concentra o maior número de nascimentos, com 189 docentes (24,67%). Os anos de 1965 (29 docentes, 3,79%) e 1964 (25 docentes, 3,26%) se destacam nesse intervalo.

No 6º período (1970–1979), foram registrados 173 docentes (22,57%), com destaque para os anos de 1978 (25 docentes, 3,26%) e 1970 e 1974 (22 docentes cada, 2,87%). O 7º

período (1980–1989) apresenta 121 registros (15,80%), sendo os anos de 1982 (22 docentes, 2,87%) e 1980 e 1981 (20 docentes cada, 2,61%) os mais representativos.

Por fim, o 8º período (1990–1994) reúne apenas 4 registros (0,52%), demonstrando a presença incipiente de docentes nascidos nos anos 1990, o que pode indicar um processo ainda inicial de renovação geracional nos programas da área.

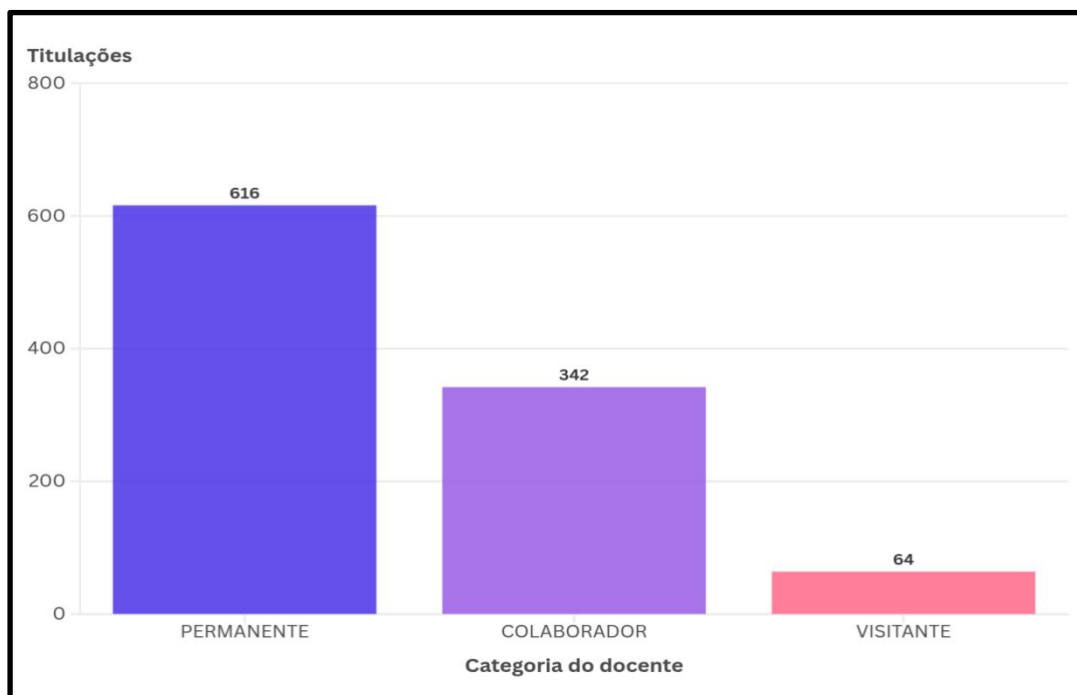
A distribuição etária dos docentes evidencia uma predominância de profissionais nascidos entre as décadas de 1950 e 1970. A expressiva presença de docentes nascidos entre 1960 e 1979 — responsáveis por aproximadamente 47,24% do total — sugere uma composição docente madura, possivelmente com larga trajetória institucional.

Por outro lado, a presença reduzida de docentes nascidos a partir da década de 1990 (apenas 0,52%) sinaliza que o processo de renovação geracional ainda é incipiente. Esse dado levanta questões sobre políticas de contratação e rejuvenescimento dos quadros, especialmente diante dos desafios contemporâneos da educação superior, como aposentadorias e a necessidade de incorporação de novas competências formativas.

Dessa forma, o perfil etário dos docentes da Ciência da Informação aponta para um quadro consolidado, porém, com potencial risco de envelhecimento, o que pode exigir atenção estratégica das instituições e das agências de fomento no planejamento da formação e inserção de novas gerações na docência e na pesquisa *stricto sensu* no país.

Dando continuidade à caracterização do corpo docente, o Gráfico 77 apresenta a distribuição das categorias de vínculo institucional dos docentes vinculados aos programas de pós-graduação em Ciência da Informação. A categorização contempla três tipos principais: permanente, colaborador e visitante, permitindo observar o grau de estabilidade e o nível de participação desses docentes nas atividades acadêmicas dos programas. É importante destacar que cada programa possui autonomia para definir a proporção, em relação ao número de docentes permanentes, para as outras duas categorias, desde que em consonância com as normas da IES à qual está vinculado e com as diretrizes da CAPES.

Gráfico 77 – Distribuição das categorias (Permanente, Colaborador, Visitante) dos docentes da CI nos programas; período 2004-2022

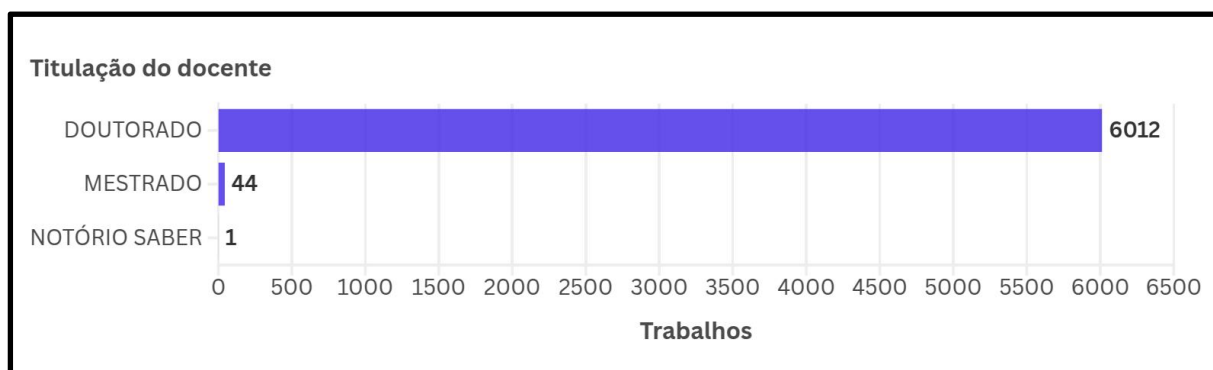


Fonte: Dados de pesquisa (2025).

Observa-se que a maioria dos registros se refere à categoria permanente, como esperado, com 616 docentes, representando 60,27% do total. Em seguida, os colaboradores somam 342 registros (33,46%), enquanto os visitantes totalizam 64 docentes, equivalendo a 6,26%. Esses dados revelam que os programas da área são majoritariamente sustentados por quadros docentes permanentes, que constituem a base das avaliações realizadas pela CAPES. A presença de docentes colaboradores e visitantes, por sua vez, indica estratégias complementares voltadas à flexibilidade institucional e ao enriquecimento acadêmico, ampliando a diversidade de experiências e expertises nos programas.

Prosseguindo com a análise do perfil docente, o Gráfico 78 apresenta a distribuição do último grau de titulação obtido pelos docentes da Ciência da Informação, revelando os níveis de formação predominantes entre os profissionais vinculados aos programas de pós-graduação da área. Esta informação contribui para compreender o alinhamento do corpo docente aos requisitos de qualificação exigidos pela CAPES para atuação na pós-graduação *stricto sensu*.

Gráfico 78 – Distribuição do último título dos docentes da CI; período 2004-2022

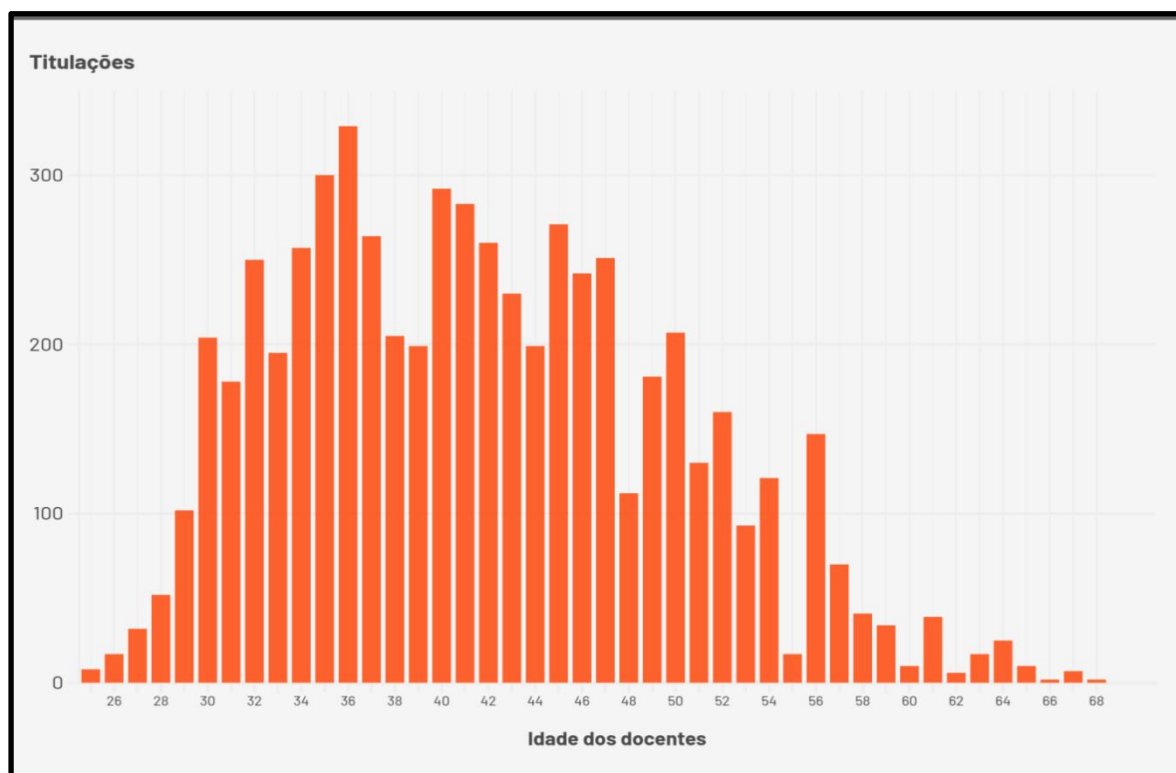


Fonte: Dados de pesquisa (2025).

Observa-se que a maioria dos docentes possui título de Doutorado, com 6.012 registros, o que representa 99,26% do total. Em seguida, nota-se um pequeno contingente de docentes com Mestrado, totalizando 44 registros (0,73%). Há ainda um caso isolado de titulação por Notório Saber, representando 0,02%. A predominância de doutores reforça o elevado nível de qualificação dos docentes que atuam nos programas da área, aspecto diretamente associado à qualidade acadêmica e aos critérios de avaliação da CAPES.

Dando continuidade à caracterização do perfil acadêmico dos docentes, o Gráfico 79 apresenta a distribuição da idade dos professores da área de Ciência da Informação no momento da obtenção de sua última titulação. Essa análise permite observar padrões etários na trajetória formativa dos docentes, identificando possíveis variações geracionais nos percursos de qualificação e contribuindo para compreender o tempo médio de inserção na pós-graduação *stricto sensu*.

Gráfico 79 – Idade do docente da CI no momento da última titulação; período 2004-2022



Fonte: Dados de pesquisa (2025).

A maior concentração de titulados encontra-se entre 30 e 45 anos, faixa etária que representa a principal janela de conclusão da formação acadêmica na área. A idade com o maior número de registros é 36 anos, com 329 docentes titulados (pico da distribuição). Em seguida, destacam-se os 35 anos (300), 34 anos (257) e 40 anos (292). De 40 a 43 anos, os valores mantêm-se elevados e relativamente estáveis, indicando que a maioria dos docentes obteve sua última titulação nesse intervalo: 41 anos (283), 42 anos (260) e 43 anos (230).

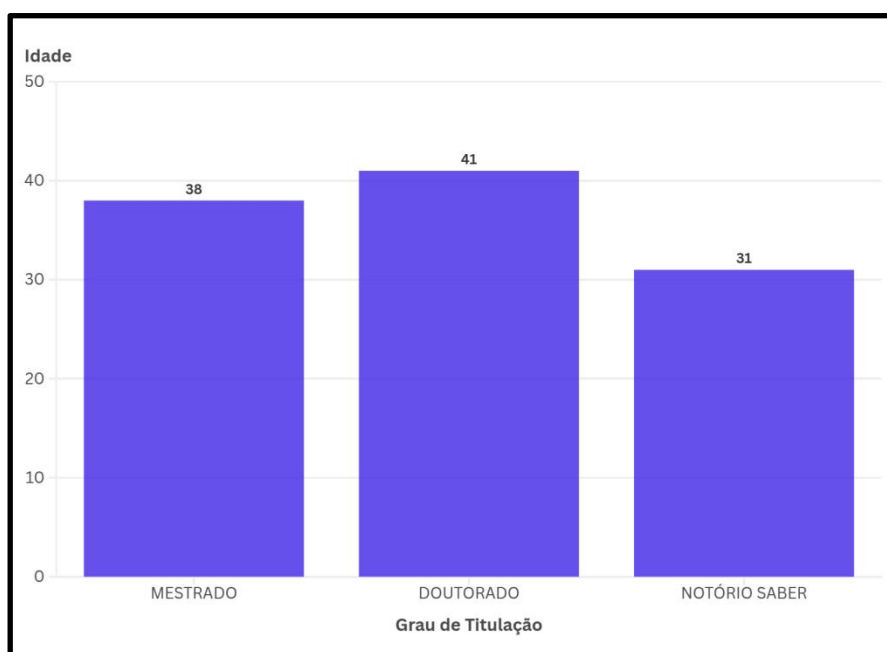
A partir dos 44 anos, observa-se uma redução progressiva na quantidade de registros. Apesar disso, idades como 45 (271), 46 (242) e 47 anos (251) ainda apresentam números expressivos. A partir dos 48 anos (112), a queda torna-se mais acentuada, com leve aumento aos 50 anos (207) e um declínio contínuo nas faixas seguintes. Aos 55 anos, por exemplo, há uma grande queda, com apenas 17 registros, com boa recuperação aos 56 anos (147). Pode-se dizer que acima dos 60 anos (10) os números são residuais, variando entre 2 e 10 docentes por faixa etária, como 61 anos (39), 62 anos (6), 66 anos (2) e 68 anos (2).

Registros de idade inferiores a 25 anos e superiores a 68 anos não foram considerados, assim como cinco registros com idade zero e um registro com idade três, por apresentarem inconsistências que podem indicar falhas de digitação ou erro no preenchimento dos dados.

Tais valores não puderam ser verificados, dado que a data de nascimento dos docentes não está disponível no banco analisado.

Complementando a análise do perfil formativo dos docentes, o Gráfico 80 apresenta a distribuição da idade média no momento da obtenção da última titulação dos professores vinculados aos programas de pós-graduação em Ciência da Informação. A comparação entre os diferentes graus de titulação permite avaliar o tempo de consolidação da trajetória acadêmica na área e sua relação com os padrões nacionais de formação.

Gráfico 80 – Distribuição da idade média por titulação dos docentes da CI; período 2004-2022



Fonte: Dados de pesquisa (2025).

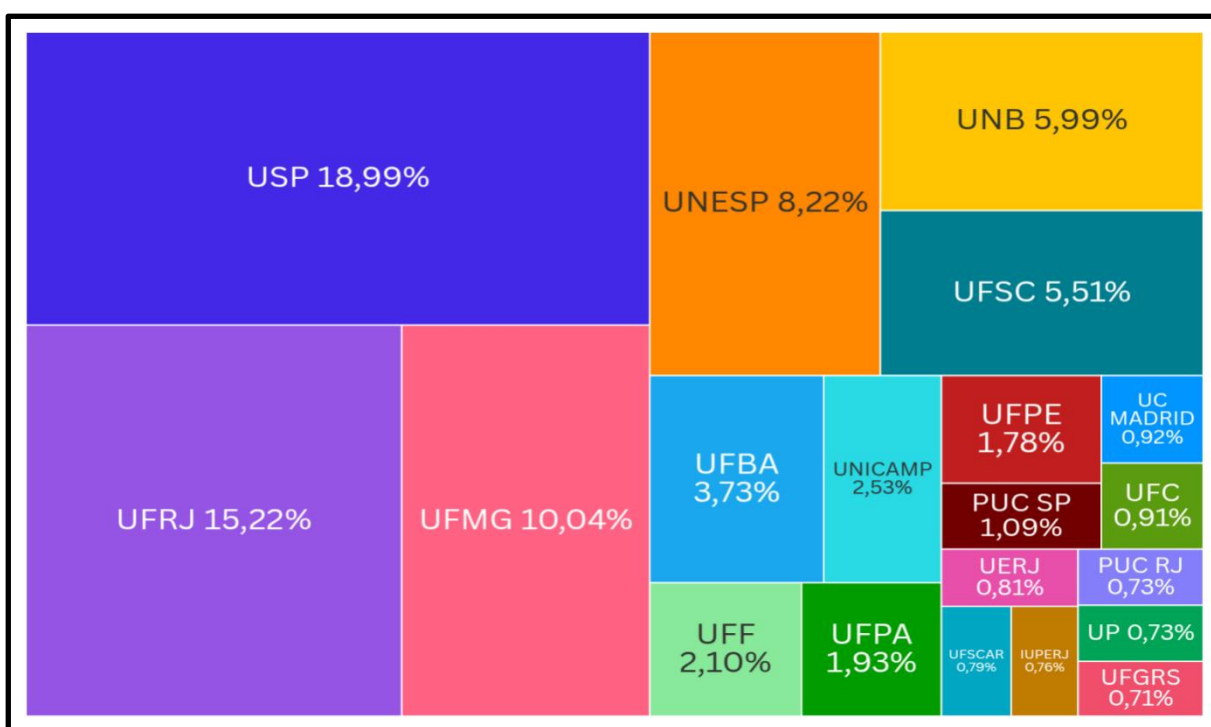
Entre os docentes analisados, a maior média de idade é observada na titulação de doutorado, com 41,68 anos, considerando 6.012 registros. Para o mestrado, a média é de 38,45 anos, com base em 44 registros. A categoria de notório saber, com apenas um docente registrado, apresenta a idade de 31 anos.

Quando comparados aos dados nacionais, os valores médios da área de Ciência da Informação se mostram superiores. Segundo o CGEE (2024), a idade média para obtenção do doutorado no Brasil variou de 39,4 anos em 1996 para 37,0 anos em 2019, com discreto aumento nos anos subsequentes. Para o mestrado, a média foi reduzida de forma consistente até atingir 32,2 anos em 2013, estabilizando-se em torno de 33,8 anos em 2021. Os dados da área sugerem que os docentes em Ciência da Informação, em geral, concluem sua formação em idades mais avançadas, o que pode estar relacionado a percursos profissionais mais longos

antes da inserção na pós-graduação ou à diversidade de formações de origem que caracterizam o campo.

Dando continuidade à caracterização do perfil formativo dos docentes, o Gráfico 81 apresenta as vinte instituições de ensino superior (IES) com maior número de titulações entre os docentes vinculados aos programas de pós-graduação em Ciência da Informação no Brasil. A visualização em *treemap* permite observar a concentração das formações em determinadas instituições, indicando polos formadores de referência no campo.

Gráfico 81 – As 20 IES com maior número de titulações entre os docentes da CI; período 2004-2022



Fonte: Dados de pesquisa (2025).

Observa-se que a USP lidera o ranking com 1.150 registros de titulação, representando 18,99% do total. Em seguida, destaca-se a UFRJ-IBICT, com 922 titulações (15,22%), e a UFMG, com 608 registros (10,04%). A UNESP aparece com 498 titulações (8,22%), seguida pela UnB, com 363 (5,99%) e a UFSC, com 334 registros (5,51%). A UFBA também figura com relevância, somando 226 registros (3,73%).

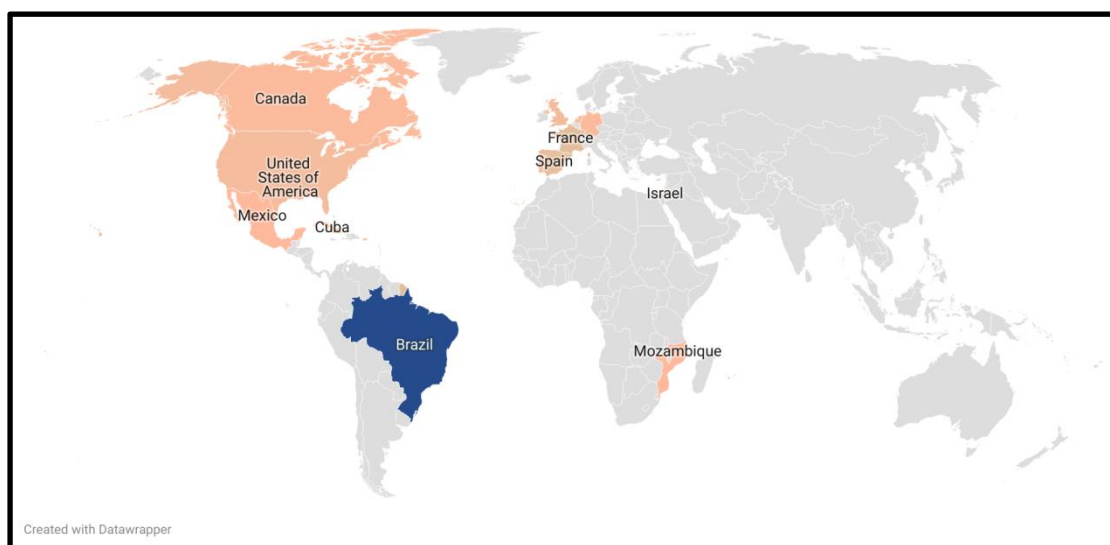
Outras instituições de destaque incluem a UNICAMP, com 153 titulações (2,53%), a UFF, com 127 (2,10%), e a UFPB, com 117 (1,93%). Entre as instituições internacionais, destaca-se a Universidad Carlos III de Madrid, com 56 registros (0,92%), acompanhada da Universidad Complutense de Madrid (38 registros – 0,63%) e da Université Paris 1 Panthéon-

Sorbonne (34 registros – 0,56%). A Universidade do Porto também aparece com 44 titulações (0,73%), mesmo número registrado pela PUC-Rio.

A presença de instituições estrangeiras indica a internacionalização parcial da formação docente na área, ao passo que a concentração em universidades públicas brasileiras evidencia o papel estruturante dessas instituições na formação de quadros docentes para a pós-graduação em Ciência da Informação.

Dando continuidade à análise do perfil formativo dos docentes da área, o Gráfico 82 apresenta a distribuição geográfica dos países onde os docentes da Ciência da Informação obtiveram sua última titulação entre os anos de 2004 e 2022. A visualização em formato de mapa mundial permite identificar os principais centros formadores, nacionais e internacionais, responsáveis pela qualificação do corpo docente da área no período analisado.

Gráfico 82 – Mapa com países da última titulação do docente da CI; período 2004-2022



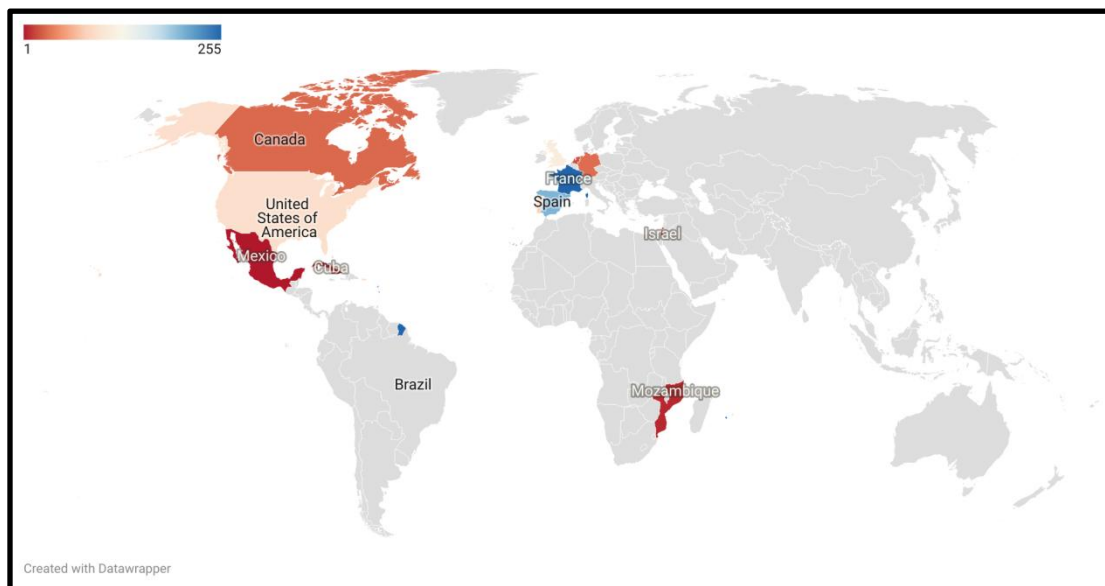
Fonte: Dados de pesquisa (2025).

Observa-se que a maioria das titulações foi realizada no Brasil, com 5.217 registros, o que representa 86,13% do total. Esse dado reforça o papel central das instituições nacionais na formação de quadros docentes da área. No entanto, observa-se também a presença de titulações realizadas no exterior, especialmente em países da Europa Ocidental e América do Norte. Essa distribuição evidencia a predominância da formação nacional, mas também indica estratégias de internacionalização dos docentes da área.

Complementando a análise da formação dos docentes, o Gráfico 83 apresenta a distribuição geográfica das titulações obtidas fora do Brasil, considerando exclusivamente os países estrangeiros onde os docentes da Ciência da Informação concluíram sua última formação no período de 2004 a 2022. A exclusão do Brasil nesta visualização permite

focalizar as dinâmicas de internacionalização da formação docente, evidenciando os principais destinos acadêmicos internacionais.

Gráfico 83 – Mapa dos países estrangeiros da última titulação dos docentes da CI (exceto Brasil); dados do período 2004-2022



Fonte: Dados de pesquisa (2025).

Observa-se que a França lidera em número de titulações realizadas no exterior, com 255 registros (4,21%), seguida pela Espanha com 204 (3,37%) e pelo Reino Unido com 114 (1,88%). Os Estados Unidos concentram 93 titulações (1,54%) e Portugal, 81 (1,34%). Esses cinco países respondem por mais de 12% das titulações obtidas no exterior, destacando-se como os principais centros formadores de docentes da área.

Outros países com presença relevante incluem a Alemanha (31 registros – 0,51%), o Canadá (29 – 0,48%) e a Holanda (20 – 0,33%). Já Israel e Moçambique aparecem com 5 registros cada (0,08%), enquanto Cuba soma 2 titulações (0,03%) e o México apenas uma (0,02%). A concentração das titulações estrangeiras em países da Europa Ocidental e América do Norte sugere uma predominância de vínculos acadêmicos com instituições de tradição acadêmica consolidada em informação, documentação e comunicação.

Dando continuidade à análise da formação docente, o Gráfico 84 apresenta a distribuição das 10 áreas básicas do conhecimento nas quais os docentes da Ciência da Informação obtiveram sua última titulação no período de 2004 a 2022. Essa visualização permite identificar os campos formativos predominantes entre os professores da pós-graduação na área, evidenciando tanto a consolidação da Ciência da Informação como área de origem quanto a diversidade interdisciplinar que a caracteriza.

Gráfico 84 – As 10 áreas básicas da última titulação dos docentes da CI com maior concentração; dados do período 2004-2022



Fonte: Dados de pesquisa (2025)²².

Observa-se que a área de Ciência da Informação lidera com 2.790 registros (46,06%), indicando que quase metade dos docentes titulou-se diretamente nesse campo. Em seguida, destacam-se áreas que tradicionalmente dialogam com a Ciência da Informação, como Comunicação (687 registros; 11,34%) e Educação (324 registros; 5,35%). História também aparece com relevância, totalizando 305 registros (5,04%), ao lado de Engenharia de Produção, com 242 registros (4,00%).

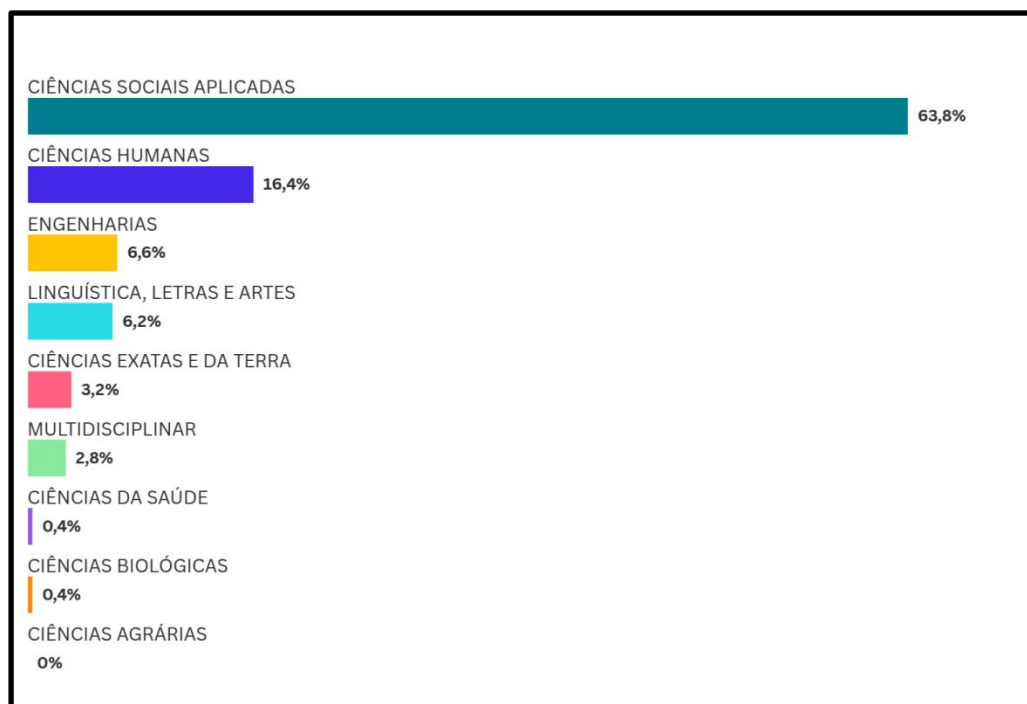
Outras áreas com representatividade são Letras (155 registros; 2,56%), Administração (139; 2,29%), Sociologia (128; 2,11%) e Ciência da Computação (123; 2,03%). Linguística completa o grupo das dez áreas mais frequentes, com 82 registros (1,35%). Essa distribuição reforça a presença de diferentes matrizes disciplinares na formação dos docentes da área, refletindo sua natureza híbrida e multifacetada.

Continuando a análise, o Gráfico 85 apresenta a distribuição das grandes áreas do conhecimento associadas à última titulação dos docentes da Ciência da Informação entre 2004 e 2022. Essa visualização permite evidenciar os campos disciplinares que formaram os

²² Fora do TOP 10 apresentado no Gráfico 85, as demais áreas básicas da última titulação dos docentes da Ciência da Informação incluem formações variadas, com destaque para Sociais e Humanidades, Engenharia Elétrica, Artes, Filosofia, Psicologia, Museologia e outras áreas aplicadas como Engenharia/Tecnologia/Gestão, totalizando percentuais individuais inferiores a 1,2%, mas que, juntas, evidenciam a diversidade e a interdisciplinaridade na formação do corpo docente da área.

docentes da área, contribuindo para o entendimento do perfil interdisciplinar em relação à classificação de áreas da CAPES.

Gráfico 85 – Distribuição das grandes áreas do conhecimento da última titulação dos docentes da CI; dados do período 2004-2022



Fonte: Dados de pesquisa (2025).

Observa-se que a maior parte das titulações está concentrada nas Ciências Sociais Aplicadas, com 3.865 registros (63,81%). Em seguida, aparecem as Ciências Humanas, com 995 registros (16,43%), seguidas pelas Engenharias, com 398 (6,57%). A área de Linguística, Letras e Artes reúne 377 registros (6,22%), enquanto a área Multidisciplinar soma 172 registros (2,84%). As Ciências Exatas e da Terra contabilizam 196 registros (3,24%), enquanto as Ciências da Saúde e as Ciências Biológicas apresentam, respectivamente, 24 (0,40%) e 25 (0,41%) registros. As Ciências Agrárias totalizam apenas 2 registros (0,03%), e há 3 registros (0,05%) classificados como "Grande Área Não Informada" (esses dois valores não aparecem no gráfico).

Encerrando a seção 5.7.7, os dados revelam um corpo docente altamente qualificado, maduro e interdisciplinar, majoritariamente vinculado a instituições públicas e formado em programas de excelência no Brasil e no exterior. A diversidade de áreas formativas, aliada à concentração em algumas universidades-chave, evidencia uma estrutura consolidada, mas também aponta desafios para a renovação geracional e a ampliação da base de formação

acadêmica. Esses elementos são cruciais para a sustentabilidade e o desenvolvimento da pós-graduação em Ciência da Informação no país.

Na sequência, a seção 5.7.8 apresenta os resultados do recorte dos dados da Ciência da Informação no DF6_discentes_CI, abordando aspectos relacionados ao perfil dos discentes vinculados aos programas, no período de 2004 a 2022.

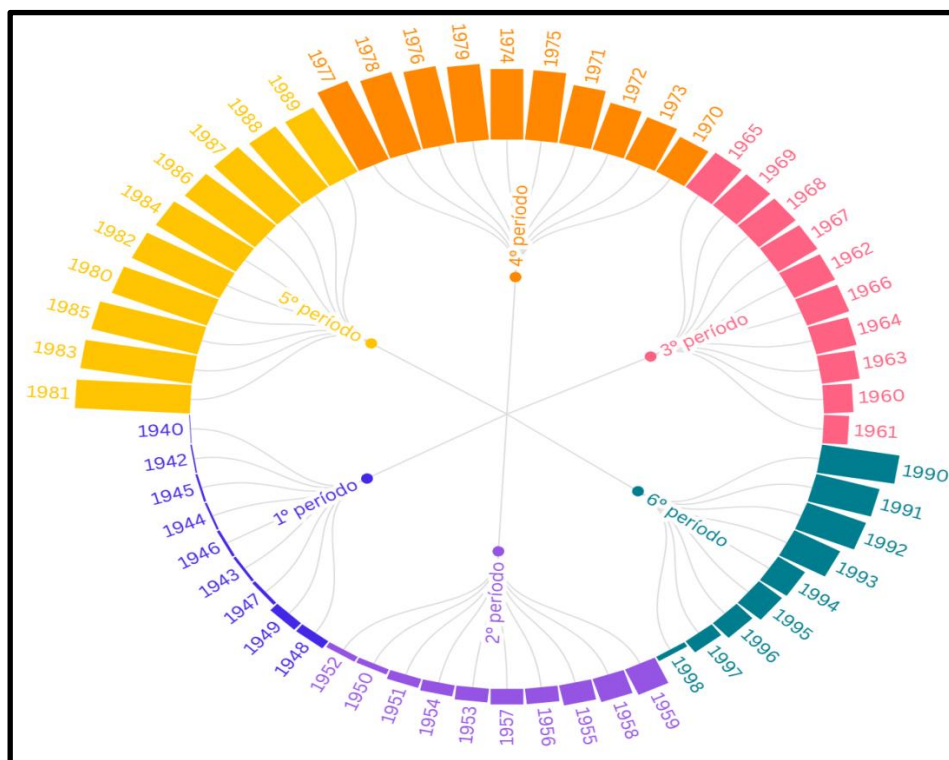
5.7.8 Resultados da etapa do Recorte dos dados da Ciência da Informação no DF6_discentes_CI

Esta seção apresenta a análise dos dados relativos aos discentes da pós-graduação *stricto sensu* na área da Ciência da Informação, com base em um total de 4.932 registros. As variáveis examinadas incluem o ano de nascimento, a nacionalidade, a faixa etária, o grau de conclusão e o tempo, em meses, necessário para a obtenção da titulação. Essas informações permitem a realização de análises descritivas e comparativas, tanto isoladas quanto combinadas, possibilitando identificar tendências, padrões formativos e variações no perfil discente ao longo do período de 2004 a 2022.

Além de caracterizar a trajetória dos estudantes, essa etapa contribui para o entendimento das dinâmicas de entrada, permanência e conclusão nos programas, oferecendo subsídios importantes para reflexões sobre políticas de formação, permanência e avaliação da pós-graduação.

Iniciando a análise discente, o Gráfico 86 apresenta a distribuição dos anos de nascimento dos estudantes da pós-graduação *stricto sensu* em Ciência da Informação no período analisado. Os dados estão organizados em seis faixas temporais (ou períodos de nascimento), a saber: 1º período: 1940–1949, 2º período: 1950–1959, 3º período: 1960–1969, 4º período: 1970–1979, 5º período: 1980–1989 e 6º período: 1990–1998. Essas faixas organizam os dados por décadas (com exceção da última, que vai até 1998, pois não há registros significativos de discentes nascidos após esse ano no conjunto analisado), facilitando a identificação de padrões geracionais ao longo do tempo.

Gráfico 86 – Distribuição dos anos de nascimento dos discentes da CI; dados do período 2004-2022



Fonte: Dados de pesquisa (2025).

Maior concentração de registros encontra-se no quinto período (1980–1989), evidenciando a predominância de discentes nascidos nessa década. Entre os anos mais representativos desse intervalo destacam-se 1981, com 160 registros (4,19%), seguido por 1983 (157 registros, 4,11%), 1985 (153 registros, 4,01%) e 1980 (142 registros, 3,72%). Outros anos com presença expressiva incluem 1982 (138 registros, 3,61%) e 1984 (134 registros, 3,51%), revelando uma concentração de nascimentos entre o início e a metade da década de 1980.

O quarto período (1970–1979) também apresenta elevado número de registros, especialmente no ano de 1977, com 138 registros (3,61%), além de 1976 (126 registros, 3,30%), 1978 (131 registros, 3,43%) e 1979 (122 registros, 3,19%), consolidando a importância dessa geração para a formação discente da área.

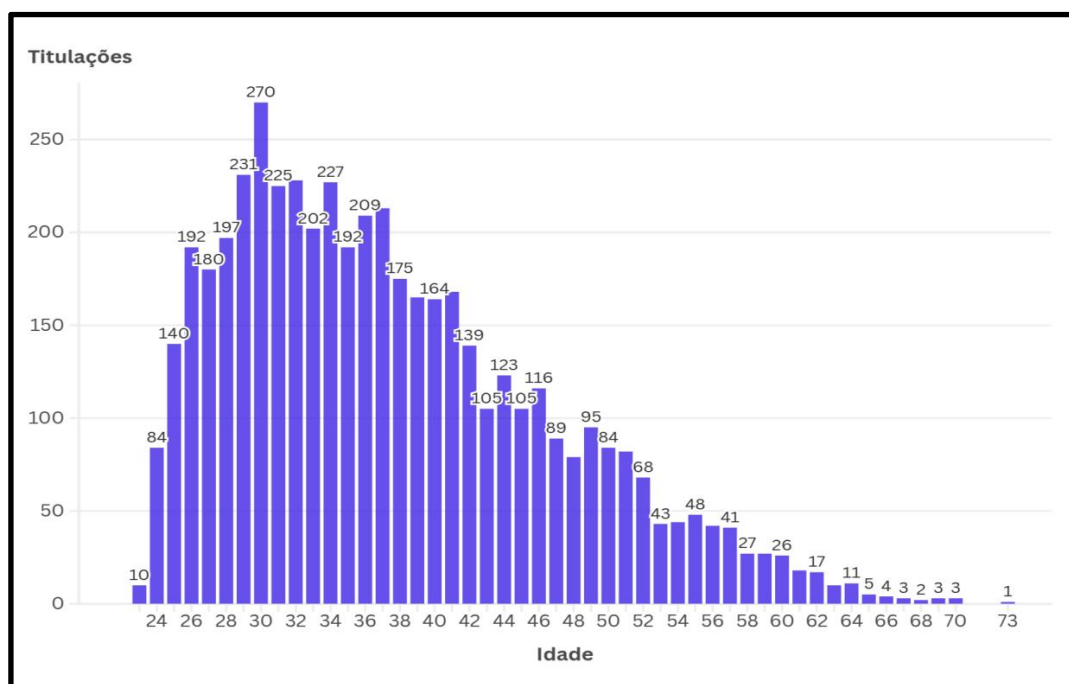
Já no sexto período (1990–1998), observa-se uma tendência de declínio gradual no número de registros. O ano de 1990 apresenta o maior volume dentro da faixa, com 109 discentes titulados (2,85%), seguido por 1991 (91 registros, 2,38%) e 1992 (88 registros, 2,30%). A partir de 1993, a frequência segue em queda, culminando com apenas 8 registros em 1998 (0,21%).

Os períodos anteriores a 1970 apresentam menor representatividade. O terceiro período (1960–1969) registra frequência decrescente à medida que os anos retrocedem, enquanto o segundo período (1950–1959) mostra maior dispersão, com destaque para 1959 (45 registros, 1,18%), 1958 (39 registros, 1,02%) e 1955 (33 registros, 0,86%). Já o primeiro período (1940–1949) possui a menor expressividade, com registros esparsos e concentrados principalmente em 1948 e 1949, ambos com 14 registros (0,37%).

Esses dados revelam que a maior parte dos discentes titulados na pós-graduação em Ciência da Informação pertence às gerações nascidas entre as décadas de 1970 e 1980, indicando um perfil etário predominantemente maduro e estável, com declínio nos registros para os nascidos a partir da década de 1990.

Prosseguindo com a análise, o Gráfico 87 apresenta a distribuição da idade dos discentes no momento da titulação na pós-graduação em Ciência da Informação, considerando a diferença entre o ano de nascimento e o ano de conclusão do curso.

Gráfico 87 – Distribuição da idade do discente no momento da titulação da CI; dados do período 2004-2022



Fonte: Dados de pesquisa (2025).

Observa-se uma maior concentração de titulações ocorre entre 26 e 40 anos, faixa etária que reúne a maior parte dos registros. A idade maior frequência no momento da titulação é 30 anos, com 270 discentes titulados (5,47%), seguida por 29 anos (231 registros, 4,68%), 32 anos (228 registros, 4,62%), 31 anos (225 registros, 4,56%) e 34 anos (227 registros, 4,60%). Outras idades com destaque incluem 28 anos (197 registros, 3,99%), 27

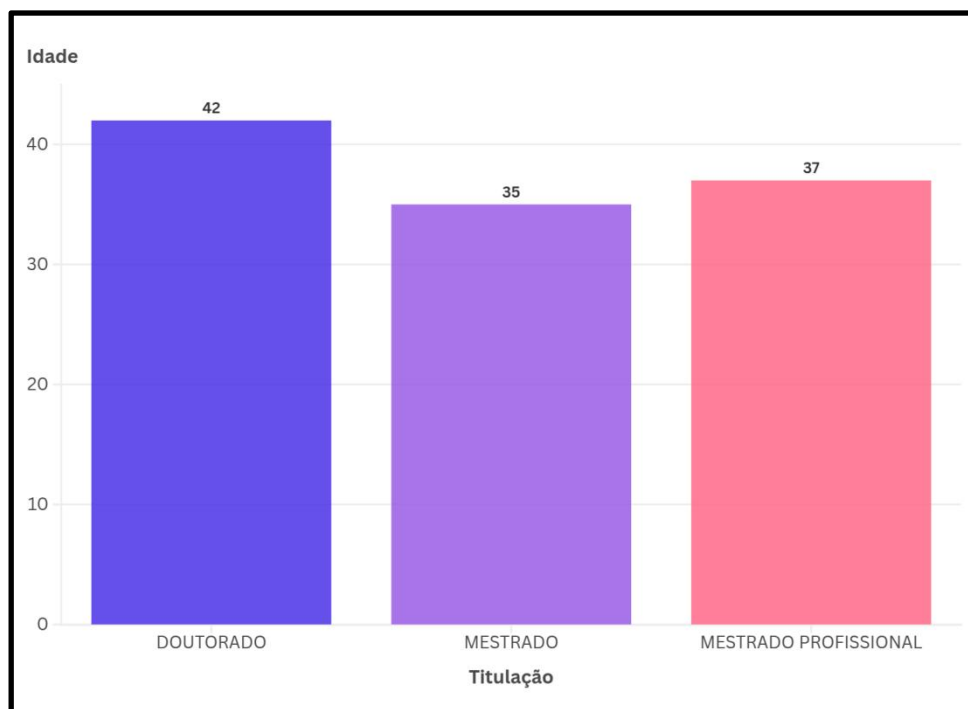
anos (180 registros, 3,65%), 26 anos (192 registros, 3,89%) e 35 anos (192 registros, 3,89%). Titulações antes dos 26 anos são menos frequentes, com 140 registros aos 25 anos (2,84%), 84 aos 24 anos (1,70%) e 10 aos 23 anos (0,20%).

A partir dos 36 anos, observa-se uma tendência de queda gradual na frequência de titulações. Ainda assim, valores expressivos aparecem aos 36 anos (209 registros, 4,24%) e 37 anos (213 registros, 4,32%), com queda subsequente: 175 registros aos 38 anos (3,55%), 165 aos 39 anos (3,35%) e 164 aos 40 anos (3,33%). Entre os 41 e 50 anos, os dados seguem em declínio: 139 registros aos 42 anos (2,82%), 123 aos 44 anos (2,49%), 105 aos 45 e 46 anos (2,13%), e 84 aos 50 anos (1,70%). Titulações em idades mais avançadas são menos frequentes, destacando-se 68 registros aos 52 anos (1,38%), 43 aos 54 anos (0,87%), 27 aos 60 anos (0,55%), 5 aos 65 anos (0,10%), 3 aos 69 anos (0,06%) e um único registro aos 73 anos (0,02%).

Essa distribuição dos dados sugere uma predominância de titulações entre os 26 e 40 anos, faixa etária que concentra os maiores volumes de conclusão dos cursos de pós-graduação em Ciência da Informação, seguida por uma redução gradual nos anos subsequentes.

Dando continuidade com a análise, o Gráfico 88 apresenta a distribuição da idade dos discentes no momento da titulação na pós-graduação em Ciência da Informação, segmentada pelos diferentes níveis de conclusão.

Gráfico 88 – Idade do docente no momento da última titulação da CI; dados do período 2004-2022



Fonte: Dados de pesquisa (2025).

A análise mostra que a titulação no mestrado acadêmico ocorre, predominantemente, entre os 30 e 36 anos, com picos aos 30 anos (203 registros), 26 anos (176 registros) e 29 anos (162 registros). As idades variam entre 23 e 69 anos, evidenciando um padrão de conclusão em idades relativamente jovens.

No caso do mestrado profissional, observa-se uma faixa etária de conclusão mais distribuída, entre 24 e 73 anos, com maior concentração entre 29 e 41 anos, destacando-se os picos aos 32 anos (52 registros), 29 anos (41 registros) e 35 e 36 anos (40 registros cada).

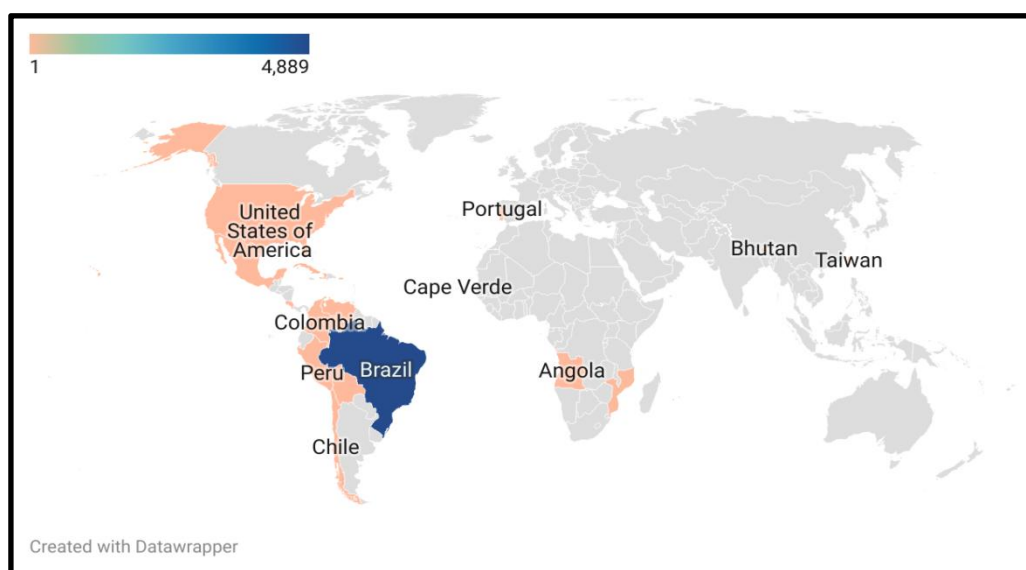
Para o doutorado acadêmico, os dados revelam um intervalo de idades mais amplo, com conclusão entre 25 e 70 anos. A maior concentração de titulações se dá entre os 29 e 42 anos, com destaque para os 39 anos (51 registros), 46 anos (50 registros) e 32 anos (48 registros). Em idades mais avançadas, há 30 registros aos 50 anos e 38 aos 51 anos, com registros pontuais de titulação após os 60 anos — como 10 aos 64 anos, 3 aos 67 anos e 1 aos 70 anos. A média de idade no momento da titulação para o doutorado acadêmico é de 42 anos, sendo superior às médias observadas para o mestrado acadêmico (35 anos) e o mestrado profissional (37 anos).

Esses dados evidenciam padrões distintos entre os níveis formativos: enquanto o mestrado acadêmico concentra-se em idades mais precoces, o mestrado profissional apresenta uma distribuição mais flexível, compatível com trajetórias acadêmicas e profissionais

simultâneas, e o doutorado acadêmico concentra um volume expressivo de titulações em idades mais avançadas.

Dando sequência à caracterização do perfil discente, o Gráfico 89 apresenta a distribuição da nacionalidade dos discentes, considerando tanto os registros de estudantes brasileiros quanto estrangeiros no período analisado.

Gráfico 89 – Mapa geral com a nacionalidade dos discentes sem o Brasil; dados do período 2004-2022

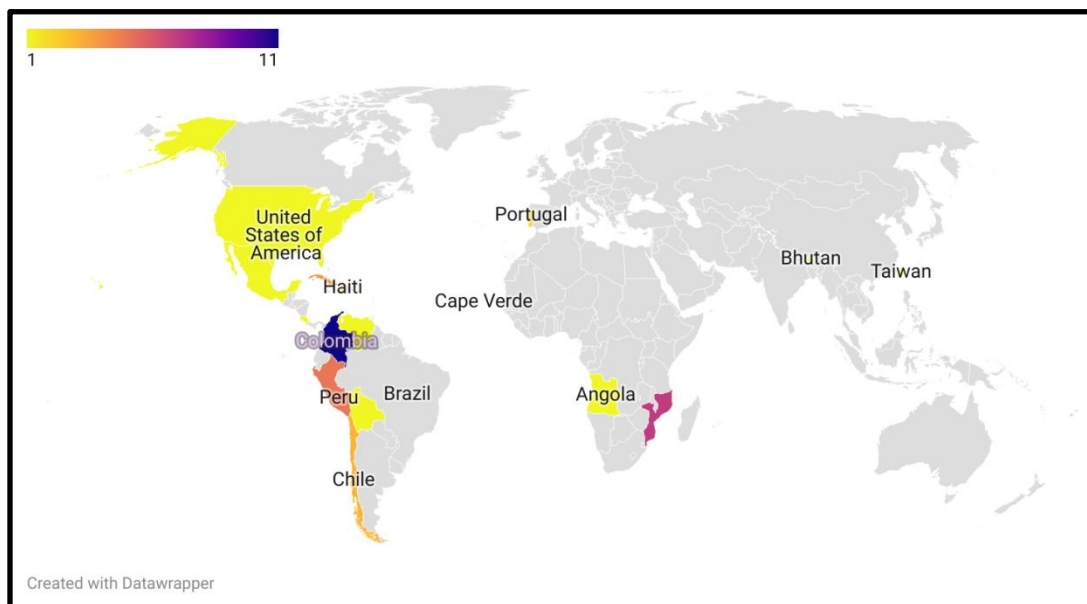


Fonte: Dados de pesquisa (2025).

Observa-se que a maioria dos registros é composta por discentes brasileiros, com 4.889 registros, representando 99,13% do total. Os discentes estrangeiros correspondem a uma parcela pequena, com 43 registros (0,87%), provenientes de diferentes continentes. Esses dados indicam uma presença limitada de estudantes internacionais nos programas da área.

Detalhando a participação estrangeira na área, o Gráfico 90 aprofunda essa análise ao apresentar exclusivamente a distribuição dos discentes estrangeiros da pós-graduação em Ciência da Informação, excluindo os registros de nacionalidade brasileira.

Gráfico 90 – Mapa com a nacionalidade dos discentes da CI sem o Brasil; dados do período 2004-2022



Fonte: Dados de pesquisa (2025).

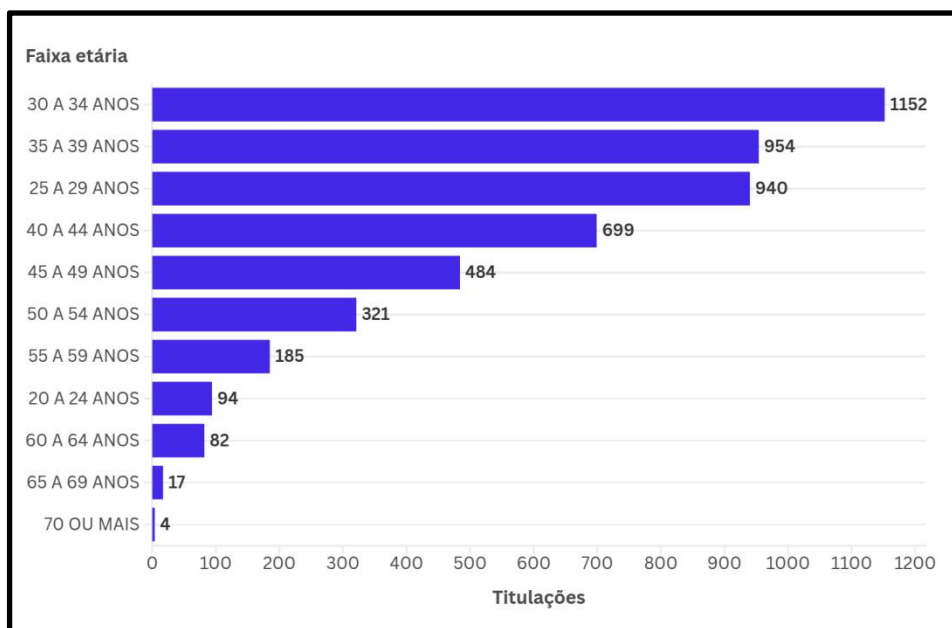
Observa-se que a América do Sul concentra a maior parte dos registros estrangeiros, com destaque para a Colômbia, que contabiliza 11 discentes (0,22%), seguida por Peru (5 registros; 0,10%) e Chile (3 registros; 0,06%). Países com menor frequência incluem Venezuela e Bolívia, com 1 registro cada (0,02%).

Na América Central e Caribe, há registros de discentes de Cuba (4; 0,08%), Haiti (2; 0,04%) e Costa Rica (1; 0,02%). A América do Norte é representada por Estados Unidos e México, com 1 registro cada (0,02%). No continente africano, destacam-se Moçambique, com 7 registros (0,14%), além de Angola e Cabo Verde, com 1 discente cada (0,02%). A Europa aparece com Portugal, que contabiliza 2 registros (0,04%). No continente asiático, figuram Taiwan e Butão, ambos com 1 registro (0,02%).

Essa distribuição revela uma presença estrangeira pontual e heterogênea nos programas de pós-graduação em Ciência da Informação no Brasil, com destaque para países da América Latina e da Comunidade dos Países de Língua Portuguesa (CPLP).

Prosseguindo com a análise do perfil dos discentes, o Gráfico 91 apresenta a distribuição da faixa etária dos discentes da pós-graduação em Ciência da Informação no período analisado, oferecendo uma visão consolidada dos perfis etários predominantes no momento da titulação.

Gráfico 91 – Faixa etária do discente da pós-graduação da CI, no momento da titulação, em termos absolutos; dados do período 2004-2022



Fonte: Dados de pesquisa (2025).

Observa-se que a maior concentração de titulações ocorre na faixa de 30 a 34 anos, com 1.152 registros, o que representa 23,36% do total. Em seguida, destaca-se a faixa de 35 a 39 anos, com 954 discentes (19,34%), muito próxima à de 25 a 29 anos, que totaliza 940 registros (19,06%). Essas três faixas etárias reúnem mais de 60% dos titulados, evidenciando um padrão de conclusão predominante entre os 25 e 39 anos.

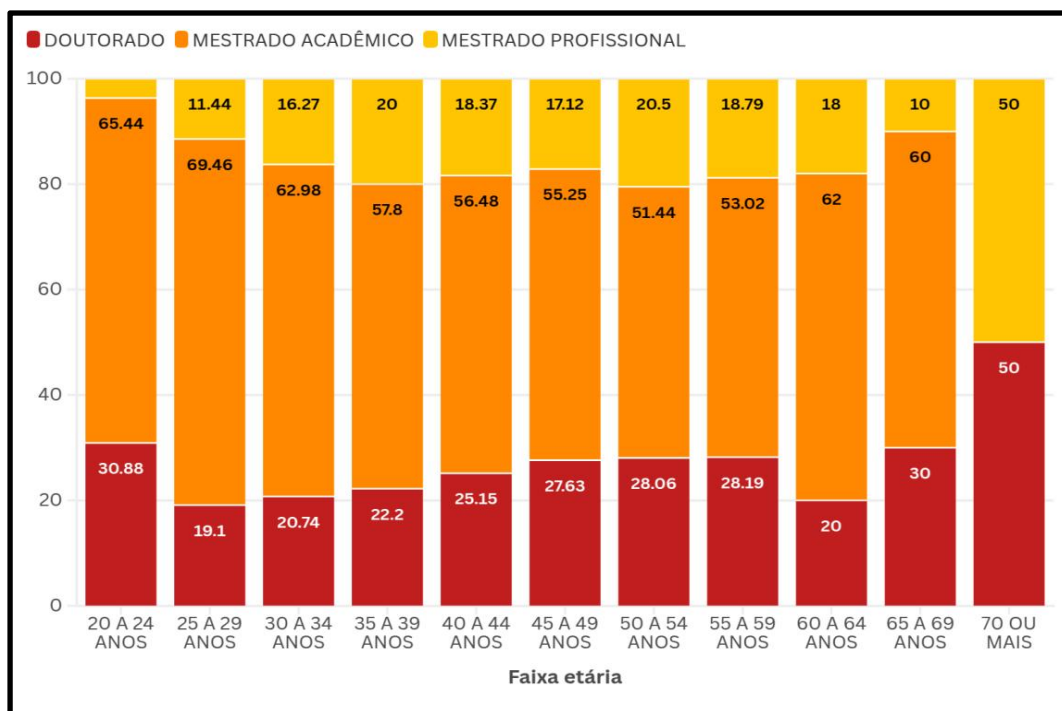
A faixa de 40 a 44 anos concentra 699 registros (14,17%), enquanto a de 45 a 49 anos contabiliza 484 discentes (9,81%). As faixas posteriores apresentam quedas mais acentuadas: 321 registros (6,51%) entre 50 e 54 anos; 185 (3,75%) entre 55 e 59 anos; e 82 registros (1,66%) entre 60 e 64 anos.

Já as faixas de 65 a 69 anos (17 registros; 0,34%) e de 70 anos ou mais (4 registros; 0,08%) representam ocorrências excepcionais, indicando que titulações em idades mais avançadas são possíveis, embora raras. E a faixa de 20 a 24 anos, com 94 registros (1,91%), representa o menor percentual entre os titulados, refletindo a baixa incidência de titulação em idade muito jovem.

Dando continuidade à análise, o Gráfico 93 aprofunda a caracterização dos discentes ao apresentar a distribuição percentual por faixa etária e por nível de titulação — mestrado acadêmico, mestrado profissional e doutorado acadêmico. Essa abordagem permite identificar a predominância de determinadas modalidades de formação em faixas etárias específicas,

evidenciando diferentes trajetórias e momentos de ingresso na pós-graduação *stricto sensu* na área da Ciência da Informação.

Gráfico 92 – Comparação faixa etária discente doutorado acadêmico, mestrado acadêmico e mestrado profissional; dados do período 2004-2022



Fonte: Dados de pesquisa (2025)²³.

A análise com foco na composição relativa por faixa etária e titulação evidencia que o mestrado acadêmico predomina nas faixas etárias mais jovens, especialmente entre 25 a 29 anos, onde representa 69,46% dos registros, e entre 20 a 24 anos, com 65,44%. A partir da faixa de 30 a 34 anos, observa-se uma redução gradual na participação relativa dessa modalidade.

O mestrado profissional apresenta uma distribuição etária distinta, com maior representatividade entre 35 e 49 anos, especialmente nas faixas de 35–39 anos (25,16%), 40–44 anos (27,18%) e 45–49 anos (30,37%). Sua participação é menos expressiva entre os mais jovens, como nas faixas de 20–24 anos (11,56%) e 25–29 anos (17,34%), o que indica que essa modalidade é majoritariamente buscada por profissionais com trajetória consolidada. A presença consistente acima dos 50 anos reforça seu papel estratégico na qualificação de perfis mais maduros, favorecendo a articulação entre experiência prática e formação acadêmica.

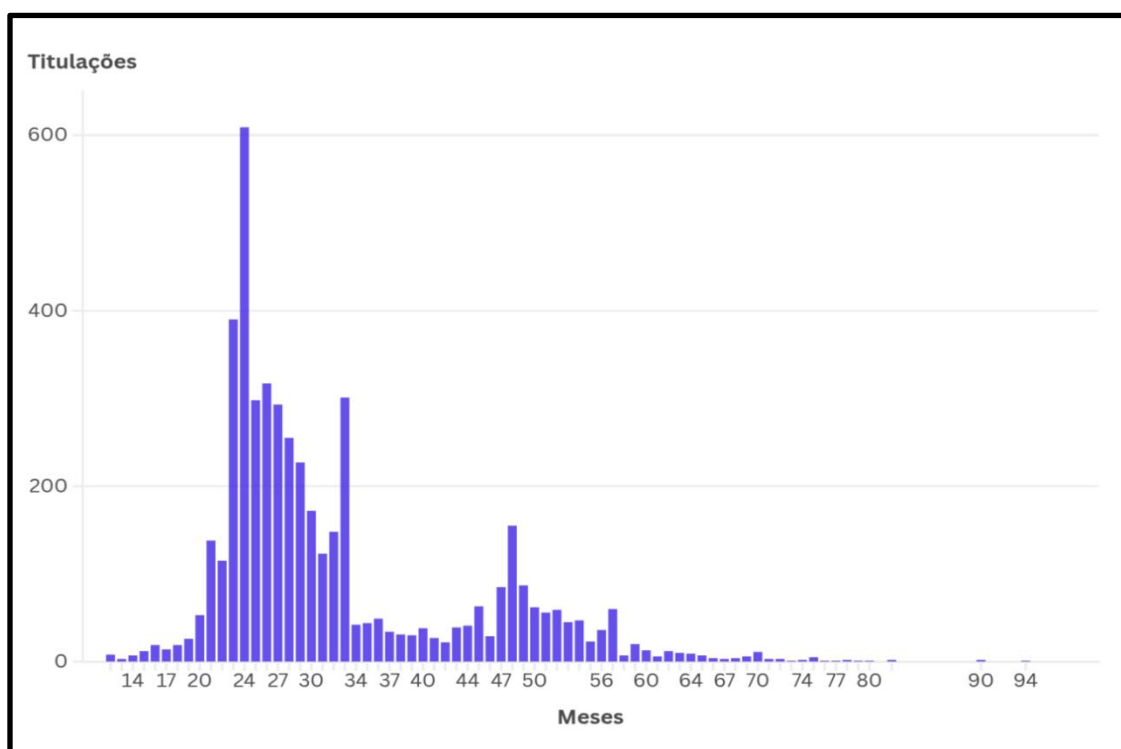
²³ Nota metodológica: O Gráfico 92 apresenta dados em percentual relativo dentro de cada faixa etária, considerando a proporção entre os três níveis de titulação em relação ao total de discentes titulados naquela faixa. Assim, os valores expressam a composição percentual de cada modalidade formativa por intervalo etário, e não o número absoluto de titulados (que está no Gráfico 91). Essa abordagem permite visualizar a predominância relativa de cada tipo de titulação conforme a idade dos discentes.

O doutorado acadêmico apresenta crescimento relativo nas faixas etárias mais elevadas, iniciando com 19,1% entre 25 e 29 anos e ultrapassando os 25% a partir da faixa de 40 anos. Esse crescimento torna-se ainda mais expressivo entre 45 e 64 anos, com o doutorado superando 28% em diversas faixas. Na faixa de 70 anos ou mais, o doutorado representa 50% dos registros, sinalizando a continuidade de trajetórias acadêmicas mesmo em estágios avançados da vida profissional.

Esses dados revelam padrões distintos entre as modalidades de titulação: o mestrado acadêmico concentra-se nas etapas iniciais da vida adulta; o profissional apresenta uma distribuição mais heterogênea, com predomínio em faixas intermediárias e avançadas; e o doutorado se concentra sobretudo em faixas etárias mais elevadas, refletindo sua exigência em termos de formação prévia e maturidade acadêmica.

Prosseguindo com a análise, o Gráfico 93 apresenta a distribuição do tempo, em meses, necessário para a titulação dos discentes da pós-graduação em Ciência da Informação no período analisado. Considerando os parâmetros estabelecidos pela Resolução nº 5, de 10 de março de 1983, que fixa prazos mínimos de 12 meses para o mestrado e 24 meses para o doutorado, foram excluídos 43 registros com tempo inferior a 12 meses e 1 registro com tempo superior a 120 meses.

Gráfico 93 – Distribuição da quantidade do tempo (meses) para titulação da CI; dados do período 2004-2022



Fonte: Dados de pesquisa (2025).

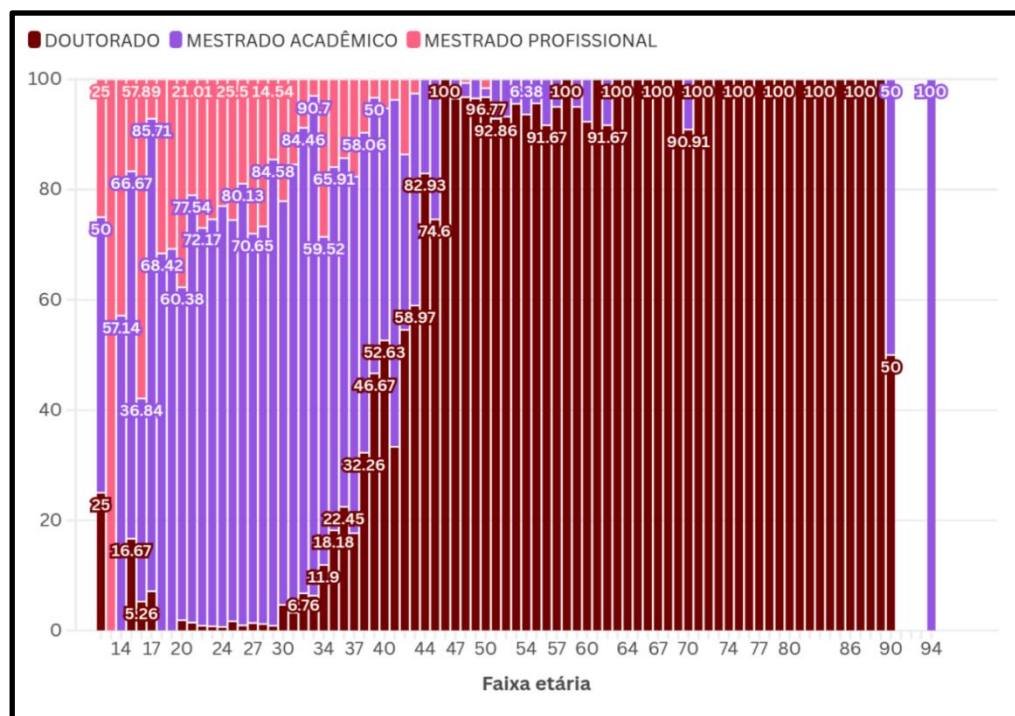
A maior concentração de titulações ocorre entre 23 e 24 meses, com 390 registros (7,91%) e 609 registros (12,35%), respectivamente, apontando para esse intervalo como o tempo predominante de conclusão, especialmente em programas de mestrado. Faixas próximas, como 25 (6,04%), 26 (6,43%) e 27 meses (5,94%), também apresentam volumes expressivos, sugerindo variações pontuais nos tempos de titulação (provavelmente, também de mestrado). Entre 28 e 30 meses, os percentuais oscilam entre 5,17% e 3,49%, reforçando a ideia de um núcleo de conclusão em torno de dois anos, característico do mestrado.

A partir dos 36 meses, observa-se uma transição nas faixas de tempo, com crescimento no número de titulações associadas a cursos de doutorado. Entre 45 e 54 meses, a concentração aumenta: há 63 registros (1,28%) com 45 meses, 85 registros (1,72%) com 47 meses e 155 registros (3,14%) com 48 meses, sinalizando que o quarto ano é o marco mais frequente de conclusão para esse nível. Faixas subsequentes, como de 49 a 54 meses, apresentam variações entre 1,76% e 0,95%.

Após os 60 meses, há uma redução acentuada no volume de titulações, com registros dispersos em intervalos mais longos, sugerindo prorrogações ou situações excepcionais. Titulações com até 94 meses ainda foram observadas, representando 0,02% do total, enquanto um caso isolado de 120 meses também foi identificado, indicando grande extensão do tempo de formação.

Dando continuidade à análise da duração da formação, o Gráfico 94 detalha a distribuição do tempo, em meses, necessário para a titulação dos discentes da pós-graduação em Ciência da Informação, segmentada por nível de titulação: mestrado acadêmico, mestrado profissional e doutorado acadêmico. A análise considera os prazos mínimos estabelecidos pela Resolução nº 5, de 10 de março de 1983 — 12 meses para o mestrado e 24 meses para o doutorado.

Gráfico 94 – Comparação meses para titulação dos discentes da CI; dados do período 2004-2022



Fonte: Dados de pesquisa (2025).

Observa-se que a maioria das titulações no mestrado acadêmico concentra-se nos intervalos de 23 a 24 meses, com 288 e 465 registros, respectivamente. Outros períodos próximos também registram números expressivos: 25 meses (217), 26 meses (254) e 27 meses (207), o que evidencia que a maioria dos discentes conclui seus cursos dentro dos prazos previstos ou com pequenas variações. A partir de 30 meses (126 registros), observa-se uma redução progressiva na quantidade de titulações, o que indica que prazos mais longos são menos frequentes nessa modalidade.

O mestrado profissional apresenta padrão semelhante, com pico em 24 meses (140 registros) e distribuição majoritariamente concentrada entre 23 e 27 meses. Essa convergência com o mestrado acadêmico indica a consolidação de práticas institucionais em torno de uma duração média de dois anos para ambas as modalidades.

Em relação ao doutorado, as titulações tornam-se mais expressivas a partir de 45 meses (47 registros), com concentração no intervalo entre 48 e 54 meses. O mês com maior número de titulações é o 48º, com 150 registros, seguido de 49 meses (84), 50 meses (60) e 52 meses (55), refletindo o padrão de conclusão em quatro anos, que é o tempo regular esperado para essa etapa. Após os 60 meses, a frequência de titulações diminui, embora ainda existam casos em períodos estendidos, como 70 meses (10 registros), 80 meses (1) e 90 meses (1). Um

caso extremo de 120 meses (1 registro) também foi identificado, indicando possível prorrogação excepcional no programa.

Esses dados revelam que a maioria dos discentes conclui seus cursos dentro dos prazos regulares definidos pelas normativas, com maiores variações especialmente no doutorado. A distribuição temporal entre as modalidades também reforça a estrutura normativa e a consolidação dos tempos médios de titulação nos programas da área.

É preciso destacar que o tempo médio de titulação é um dos critérios que impactam diretamente a nota dos programas de pós-graduação no Formulário de Avaliação Quadrienal da CAPES. Para programas com conceitos 5, 6 ou 7, espera-se que a maioria dos estudantes conclua seus trabalhos dentro do prazo regulamentar ou em períodos próximos — no caso do doutorado, até 60 meses. A extensão desses prazos, como defesas realizadas entre 90 e 120 meses (7,5 a 10 anos), pode ser interpretada como indício de desorganização do programa ou de falhas no acompanhamento dos discentes, comprometendo a avaliação institucional.

Finalizada a análise dos dados referentes aos discentes da pós-graduação *stricto sensu* em Ciência da Informação, apresentada nesta seção, foi possível traçar um panorama sobre o perfil formativo e sociodemográfico da comunidade estudantil da área entre 2004 e 2022. Observou-se uma predominância de discentes brasileiros. Em relação à faixa etária, identificou-se que a maioria das titulações ocorre entre os 30 e 39 anos, com variações específicas conforme o tipo de curso, revelando diferentes dinâmicas entre o mestrado acadêmico, o profissional e o doutorado.

Os dados também indicam padrões de titulação consistentes com os prazos regulamentares: o mestrado, em suas duas modalidades, apresenta tempo de conclusão concentrado entre 23 e 27 meses; já o doutorado, entre 45 e 54 meses. A idade média e o tempo de formação reforçam a existência de trajetórias acadêmicas diversificadas, o que reflete a maturidade e a pluralidade da área em termos de formação e atuação.

Essa caracterização quantitativa dos discentes fornece subsídios relevantes para a compreensão da estrutura da pós-graduação em Ciência da Informação no Brasil, ao mesmo tempo em que contribui para reflexões sobre políticas de formação, acesso, permanência e internacionalização.

Com base nas informações sistematizadas até aqui, o próximo capítulo avança na Análise de Resultados, aprofundando a investigação sobre os dados e relações reveladas pelas variáveis, explorando recortes qualitativos e interpretações críticas que visam compreender os fatores estruturantes da formação, atuação e consolidação da área de Ciência da Informação no país.

6 ANÁLISE DOS RESULTADOS

Este capítulo analisa criticamente os achados do Capítulo 5, identificando padrões, tendências e desafios da pós-graduação em Ciência da Informação no Brasil entre 2004 e 2022. A partir da estruturação dos dados e das metodologias aplicadas, busca-se interpretar os resultados, evidenciando semelhanças, contrastes e particularidades da área no contexto da pós-graduação nacional.

A abordagem exploratória permite mapear regularidades e variações nos indicadores acadêmicos, analisando a evolução da produção científica, dos programas, docentes e discentes. A comparação com a pós-graduação brasileira como um todo destaca especificidades da Ciência da Informação, relacionando sua trajetória a aspectos institucionais, temáticos e metodológicos. A discussão dos resultados é conduzida à luz da literatura e das políticas institucionais que impactam a pós-graduação, incluindo diretrizes da CAPES e políticas científicas. Ao articular evidências empíricas com referenciais normativos, busca-se compreender a dinâmica da área e os fatores que influenciam sua produção acadêmica.

Para tanto, a análise é organizada em seis seções. A seção 6.1 discute a qualidade dos dados, com ênfase nas discrepâncias e anomalias encontradas durante a curadoria. A seção 6.2 apresenta a interpretação dos principais resultados referentes à Ciência da Informação, enquanto a seção 6.3 estabelece um diálogo com estudos anteriores da área, contribuindo para a consolidação de um panorama comparativo. A seção 6.4 aprofunda a análise da atuação da Ciência da Informação na pós-graduação brasileira. Na 6.5, são sistematizados os principais achados do capítulo, e, por fim, a seção 6.6 propõe uma metodologia que sintetiza o percurso desenvolvido e oferece subsídios para futuras investigações no campo.

6.1 Qualidade dos dados: Discrepâncias e Anomalias

A análise dos dados da pós-graduação em Ciência da Informação entre 2004 e 2022 revelou diferentes tipos de inconsistências e variações que afetam diretamente a confiabilidade, a comparabilidade e a interpretação dos resultados. Um dos primeiros aspectos observados foi a distribuição geográfica desigual dos programas de pós-graduação. A maior parte encontra-se concentrada nas regiões Sudeste e Sul, reflexo do histórico de investimento em infraestrutura acadêmica, da consolidação de universidades de pesquisa e da maior disponibilidade de financiamento federal e estadual. Em contraste, as regiões Norte e Centro-Oeste apresentam menor presença, o que pode indicar barreiras institucionais, escassez de pesquisadores e dificuldade na captação de recursos. Essa assimetria reforça a necessidade de

políticas públicas que incentivem a interiorização da pós-graduação e a consolidação de programas em regiões historicamente menos contempladas.

O volume de teses e dissertações defendidas ao longo dos anos apresenta variações importantes. De 2004 a 2019, houve crescimento expressivo na produção, seguido por uma queda acentuada em 2020, associada ao impacto da pandemia da COVID-19. A partir de 2021, observa-se uma recuperação gradual, embora ainda aquém dos níveis anteriores. Além do contexto sanitário, outros fatores institucionais também influenciaram essa variação, como alterações nos critérios de avaliação da CAPES, mudanças na concessão de bolsas e reestruturações internas dos programas. A capacidade de orientação e o número de docentes permanentes ativos também desempenham papel central no ritmo de defesas anuais. Assim, compreender essas oscilações se torna necessário para avaliar a estabilidade da área e os fatores que impactam a continuidade das pesquisas.

Nos títulos e resumos dos trabalhos, identificou-se uma expressiva diversidade terminológica, dificultando a padronização e a análise temática longitudinal. Termos distintos são usados para descrever conceitos similares, o que compromete a recuperação da informação e a comparação entre pesquisas. Essa diversidade reflete tanto a ausência de um vocabulário controlado adotado no sistema quanto a pluralidade de abordagens teóricas e metodológicas na área. A evolução dos termos ao longo do tempo também evidencia a incorporação de novas temáticas, o que, embora positivo, exige estratégias que favoreçam maior uniformidade terminológica. A falta de padronização impacta diretamente a realização de estudos comparativos e pode dificultar a identificação de relações entre diferentes pesquisas. A adoção de diretrizes mais uniformes para a nomenclatura utilizada nos trabalhos acadêmicos poderia contribuir e facilitar análises da produção científica.

O campo relacionado à titulação e ao tempo de conclusão dos cursos revelou discrepâncias relevantes. Há registros de titulação em prazos muito inferiores ou superiores ao esperado para mestrado e doutorado, o que pode ser decorrente de diferenças entre programas, alterações curriculares, ajustes administrativos, ou mesmo de erros de registro. Casos de defesa sem titulação registrada e registros com valores nulos ou inconsistentes também foram identificados. Essa falta de padronização nos prazos e registros pode comprometer a análise da trajetória acadêmica e dificultar a comparação entre programas e instituições. E compreender essas diferenças garante a confiabilidade dos dados e para que futuras pesquisas possam interpretar corretamente as trajetórias acadêmicas dos egressos da área.

Outro ponto crítico refere-se à incompletude de dados sobre orientadores e discentes. Em muitos casos, não há registro do orientador responsável, dificultando a análise das redes

de orientação, colaborações institucionais e evolução das linhas de pesquisa. Nos dados de discentes, observam-se variações na grafia dos nomes, duplicações e ausência de identificadores únicos, o que prejudica o rastreamento acadêmico e impede a vinculação automática entre egressos (ex-discentes) e docentes ativos. Essas falhas estão ligadas à forma como os dados são inseridos nos sistemas da CAPES, frequentemente de maneira manual pelos programas. A falta de qualidade nos dados compromete estudos sobre a formação de recursos humanos, a evolução dos programas e a distribuição dos pesquisadores nos diversos campos, exigindo um tratamento de dados rigoroso durante as pesquisas.

Tais limitações indicam a necessidade urgente de melhorias na estruturação e na governança dos dados da pós-graduação pela CAPES. A adoção de identificadores únicos para docentes e discentes, associada a um modelo de dados padronizado para nomes de programas, áreas e instituições, contribuiria para a redução de erros e a melhoria na rastreabilidade dos dados. Além disso, a obrigatoriedade do preenchimento de campos críticos, como informações completas de orientação, permitiria análises mais aprofundadas e confiáveis sobre a formação de recursos humanos e a evolução dos programas.

A incorporação dos princípios FAIR (Findable, Accessible, Interoperable, Reusable) nos sistemas de dados da CAPES poderia aprimorar a transparência e a interoperabilidade das informações. A criação de metadados padronizados, a vinculação a identificadores persistentes (como ORCID e DOI) e a interoperabilidade entre sistemas aumentariam a confiabilidade dos dados e permitiriam o alinhamento com práticas internacionais. Tais iniciativas fortaleceriam a capacidade da CAPES de oferecer dados de qualidade para o planejamento de políticas públicas e para o avanço da pesquisa no país.

6.2 Interpretação dos principais resultados: a Ciência da Informação

A evolução da pós-graduação em Ciência da Informação no Brasil apresenta um crescimento contínuo ao longo dos anos, evidenciado pelo aumento na quantidade de discentes, docentes e programas acadêmicos. A análise dos dados indica uma ampliação progressiva do número de matrículas, refletindo a consolidação da área como um campo de pesquisa e formação. O crescimento do corpo discente acompanha a expansão dos programas de pós-graduação, indicando um interesse crescente pela formação avançada na área. Esse fenômeno pode estar relacionado tanto à maior demanda por profissionais especializados em gestão da informação e tecnologia quanto ao fortalecimento das políticas institucionais de incentivo à qualificação acadêmica. Além disso, a diversificação dos perfis de pesquisa e das

linhas temáticas ao longo dos anos sugere um amadurecimento da área, acompanhando as transformações tecnológicas e sociais que impactam o campo da Ciência da Informação.

O aumento do número de docentes vinculados aos programas de pós-graduação em Ciência da Informação reflete a ampliação das instituições que oferecem cursos nessa área e a necessidade de qualificação contínua do corpo acadêmico. A distribuição desses docentes, no entanto, ainda apresenta variações regionais, com maior concentração em determinadas instituições e menor presença em outras localidades. Essa expansão, embora positiva, exige uma análise sobre os desafios relacionados à distribuição da oferta de cursos e à infraestrutura disponível para o desenvolvimento das pesquisas. A ampliação dos programas também tem impactos na produção científica da área, com um crescimento no volume de dissertações e teses defendidas, sugerindo um fortalecimento do campo no cenário acadêmico nacional. Entretanto, é fundamental considerar as assimetrias no desenvolvimento dos programas, bem como as políticas de fomento que influenciam diretamente a capacidade das instituições de sustentar esse crescimento ao longo dos anos.

O crescimento da pós-graduação em Ciência da Informação no Brasil entre 2004 e 2022 é evidenciado pela expansão do número de discentes, docentes e programas acadêmicos. A análise dos dados demonstra um aumento progressivo na quantidade de titulados ao longo dos anos, com variações nos períodos de maior e menor crescimento. A ampliação da oferta de cursos, tanto em instituições públicas quanto privadas, contribuiu para esse movimento, refletindo a consolidação da área no cenário acadêmico nacional. A distribuição dos titulados por idade também apresenta um padrão consistente, com maior concentração entre 27 e 45 anos, indicando um fluxo regular de formação de pesquisadores e profissionais qualificados para atuar no campo da Ciência da Informação .

Além do crescimento no número de titulados, observa-se um aumento na diversificação dos programas de pós-graduação. A análise revela que a titulação em mestrado ocorre, em média, aos 24 anos, enquanto a obtenção do doutorado é registrada predominantemente aos 36 anos. Essa variação na idade média dos titulados reflete diferentes trajetórias acadêmicas e profissionais dentro da área. Ademais, os dados apontam para a existência de programas acadêmicos e profissionais com distintos perfis de formação, o que amplia as possibilidades de inserção dos egressos em diferentes setores do mercado e da pesquisa. A relação entre a quantidade de novos programas e a expansão da área indica um cenário de fortalecimento institucional e de maior reconhecimento da Ciência da Informação como campo relevante no desenvolvimento científico e tecnológico .

A análise da produtividade acadêmica na área da Ciência da Informação entre 2004 e 2022 revela padrões de crescimento e variações ao longo dos anos. Durante esse período, houve um aumento expressivo no número de trabalhos defendidos, especialmente entre 2004 e 2019, refletindo a ampliação da pós-graduação e a consolidação dos programas na área. Esse crescimento foi mais acentuado nos anos de 2005, 2012 e 2013, que registraram os maiores acréscimos percentuais na produção acadêmica. Entretanto, a partir de 2020, observa-se uma interrupção nesse crescimento, com uma redução no volume de trabalhos defendidos. Esse declínio pode estar relacionado a fatores institucionais e conjunturais, como a pandemia da COVID-19, que impactou diretamente a continuidade das pesquisas e a defesa de dissertações e teses.

O período de 2021 e 2022 mostra uma tentativa de retomada da produtividade acadêmica na Ciência da Informação, com um leve aumento no número de trabalhos defendidos. No entanto, essa recuperação ainda não atingiu os patamares observados em 2017, evidenciando uma estabilização em níveis inferiores ao crescimento registrado na década anterior. Essa tendência reforça a necessidade de compreender as variáveis que afetam a produção acadêmica, incluindo o impacto de políticas institucionais, financiamento da pesquisa e condições de trabalho dos discentes e docentes. Além disso, a análise sugere que fatores estruturais, como a expansão da pós-graduação profissional e as mudanças nas exigências de titulação, possivelmente exerceram influência sobre as variações no volume de produção acadêmica ao longo do período analisado.

Em 2017, a avaliação da pós-graduação passou por uma reformulação, priorizando indicadores qualitativos sem desconsiderar a relevância dos dados quantitativos. A reestruturação dos critérios avaliativos buscou garantir uma abordagem mais abrangente, conferindo peso equivalente a diferentes dimensões dos programas, como estrutura organizacional, qualidade da formação acadêmica e impacto da produção científica na sociedade. Em 2018, foi introduzida a autoavaliação como um componente da avaliação externa, incentivando maior participação da comunidade acadêmica no processo avaliativo. Esse modelo visou superar a concepção tradicional baseada exclusivamente na avaliação externa e promover uma abordagem mais formativa, permitindo que os programas identificassem suas próprias potencialidades e fragilidades.

Além das mudanças estruturais na avaliação da pós-graduação, fatores externos também impactaram a produção acadêmica no período analisado. A pandemia de COVID-19, iniciada em 2020, resultou em restrições institucionais que afetaram diretamente o número de defesas de dissertações e teses. O fechamento de laboratórios, bibliotecas e espaços de

pesquisa, aliado às limitações na realização de trabalhos de campo, gerou atrasos na conclusão de investigações acadêmicas. Em algumas instituições, verificou-se uma redução expressiva no número de titulados, refletindo as dificuldades enfrentadas pelos discentes na adaptação às novas condições impostas pelas restrições sanitárias. Esse cenário evidenciou a necessidade de maior flexibilidade nos prazos e processos de avaliação, além da implementação de estratégias que garantam a continuidade das pesquisas em contextos adversos, preservando a qualidade e a regularidade da produção científica na pós-graduação brasileira.

A análise da produção acadêmica na área de Ciência da Informação revela uma concentração de teses e dissertações em um grupo restrito de instituições de ensino superior, refletindo tanto a consolidação histórica desses programas quanto a infraestrutura disponível para pesquisa e formação na área. Entre as cinco instituições com maior volume de trabalhos defendidos, destacam-se a Universidade de São Paulo (USP), Universidade Federal de Minas Gerais (UFMG), Universidade de Brasília (UnB), Universidade Estadual Paulista (UNESP) e Universidade Federal da Paraíba (UFPB). Essas instituições possuem programas consolidados e alta produtividade acadêmica, exercendo papel central na formação de mestres e doutores na área. Além do número expressivo de trabalhos, essas universidades se destacam pela diversidade de temas abordados e pela capacidade de atração de pesquisadores, financiamentos e colaborações interinstitucionais. O elevado volume de produções acadêmicas nesses centros também pode ser atribuído à presença de programas de pós-graduação avaliados com notas altas pela CAPES, o que favorece a captação de recursos e a permanência de docentes altamente produtivos.

Além do volume de produção acadêmica, a antiguidade dos programas também tem papel relevante na estruturação da área no Brasil. Entre as três instituições pioneiras na criação de programas na Ciência da Informação, encontram-se a Universidade de Brasília (UnB), a Universidade Federal de Minas Gerais (UFMG) e a Universidade Federal do Rio de Janeiro (UFRJ-IBICT). Esses programas, estabelecidos nas décadas iniciais da consolidação da área no país, contribuíram para a definição das diretrizes curriculares e das principais linhas de pesquisa em Ciência da Informação. A interseção entre essas instituições pioneiras e as que apresentam maior número de trabalhos defendidos reforça a ideia de que a tradição acadêmica e a capacidade de inovação são fatores determinantes para a produção científica na Ciência da Informação no Brasil.

6.3 Comparação dos resultados com estudos anteriores na CI

Os dados apresentados nos resultados confirmam o crescimento expressivo da pós-graduação na área, com um aumento contínuo no número de programas, docentes e discentes ao longo do período analisado. Esse crescimento reflete o processo de institucionalização da área no Brasil, conforme indicado na revisão de literatura. Além disso, a diversificação das linhas de pesquisa observada na revisão também se manifesta nos dados empíricos, com um aumento na produção acadêmica em temas emergentes e uma maior distribuição geográfica dos programas, apesar da concentração em determinadas regiões. A revisão da literatura apontava uma crescente visibilidade da área da Ciência da Informação no cenário acadêmico nacional e internacional, associada ao aumento da produção científica e ao fortalecimento das redes de colaboração. Nos resultados, verifica-se que o número de dissertações e teses defendidas na área cresceu ao longo dos anos, corroborando a tendência identificada nos estudos bibliométricos.

Os dados apresentados nos resultados evidenciam um aumento no uso de métricas e indicadores para avaliação da produção acadêmica na Ciência da Informação. A literatura revisada já apontava essa tendência, especialmente com a adoção de ferramentas como VOSviewer e Gephi para mapear redes de colaboração e a influência crescente dos indicadores de impacto. A estruturação dos dados da CAPES utilizada no estudo reforça esse cenário, demonstrando a necessidade de abordagens quantitativas e qualitativas para compreender a evolução da pesquisa na área.

A revisão de literatura indicava uma ampliação nos temas investigados, com um aumento de pesquisas sobre ciência aberta, gestão de dados e políticas de acesso à informação. Os dados apresentados nos resultados mostram essa diversificação, com um crescimento na produção acadêmica relacionada à dados e outras abordagens interdisciplinares. Esse aspecto reforça a adaptação da área às novas demandas do mercado e da sociedade, consolidando sua natureza dinâmica e sua aproximação com diferentes campos do conhecimento.

Apesar dos avanços, tanto a revisão da literatura quanto os dados empíricos apontam desafios estruturais na pós-graduação em Ciência da Informação. A concentração dos programas nas regiões Sul e Sudeste, a desigualdade no acesso a financiamento e a falta de padronização dos dados científicos são questões que permanecem como obstáculos para um desenvolvimento mais igualitário da área. Além disso, os dados revelam dificuldades na integração entre programas e grupos de pesquisa, confirmando a necessidade de políticas mais

eficazes para promover maior colaboração e compartilhamento de informações entre as instituições.

Os estudos revisados apontam um crescimento expressivo dos programas de pós-graduação em Ciência da Informação, acompanhando a expansão da pós-graduação no Brasil desde os anos 2000. Esse crescimento está relacionado à demanda por profissionais qualificados e ao incentivo de políticas de fomento, como as da CAPES. Além do aumento no número de programas, há uma diversificação das linhas de pesquisa, refletindo o caráter multidisciplinar da área. A produção acadêmica também apresenta crescimento contínuo, com aumento no número de dissertações e teses defendidas. Esse cenário se alinha a outras áreas do conhecimento, onde a colaboração científica, o financiamento à pesquisa e o acesso a repositórios digitais têm impulsionado a produção científica. Tecnologias de gestão da informação e análise de dados têm desempenhado um papel importante na difusão do conhecimento na Ciência da Informação.

A avaliação da pós-graduação passou por mudanças nos últimos anos, com a CAPES incorporando critérios qualitativos e ampliando a análise dos impactos acadêmicos e sociais da produção científica. Estudos indicam que essa transformação não é exclusiva da Ciência da Informação, mas reflete um movimento mais amplo dentro da pós-graduação brasileira. A literatura também destaca a ampliação dos temas de pesquisa, abrangendo áreas como gestão de dados, ciência aberta, inteligência artificial e aprendizado de máquina. Esse movimento tem fortalecido a interdisciplinaridade da área, aproximando-a de outros campos.

Os estudos revisados também indicam desafios persistentes na distribuição regional dos programas de pós-graduação. A maior concentração no Sudeste e no Sul contrasta com a presença reduzida nas regiões Norte e Nordeste. A desigualdade no acesso a recursos e infraestrutura acadêmica impacta a formação de pesquisadores e a consolidação da área em nível nacional. Além disso, há dificuldades na padronização dos dados da produção científica, o que afeta estudos bibliométricos e avaliações institucionais.

A análise dos dados apresentados no capítulo 5 evidencia padrões que confirmam tendências previamente identificadas na revisão de literatura, ao mesmo tempo que destaca aspectos novos que não haviam sido explorados. A expansão da pós-graduação em Ciência da Informação, apontada na literatura como um movimento consolidado desde os anos 2000, é observada nos resultados, tanto no crescimento do número de programas de mestrado e doutorado quanto na diversificação das linhas de pesquisa. Além disso, os dados confirmam a concentração desses programas nas regiões mais desenvolvidas do país, reforçando a desigualdade regional já destacada nos estudos revisados. A relação entre o crescimento da

produção acadêmica e as mudanças nos critérios de avaliação da CAPES também se manifesta nos resultados, evidenciando uma adaptação dos programas às novas exigências institucionais.

Os achados sobre a adoção de métodos bibliométricos e o uso de novas tecnologias para mapear redes de colaboração são coerentes com o que a literatura aponta como um avanço na área. O capítulo 5 mostra um crescimento na produção de dissertações e teses, refletindo a ampliação das redes de pesquisa e a adoção de ferramentas para análise da produção científica. A interdisciplinaridade também é confirmada nos dados, indicando que a Ciência da Informação tem absorvido conceitos de outras áreas, como ciência da computação e administração, em suas investigações. No entanto, os resultados indicam que, apesar dessa ampliação, há uma variação na consolidação dessas abordagens entre os programas, sugerindo que nem todas as instituições acompanham essa diversificação no mesmo ritmo.

Além dessas confirmações, os resultados do capítulo 5 revelam novas questões que não foram abordadas na revisão de literatura. Um dos achados inéditos é a inconsistência nos registros acadêmicos, como a ausência de identificação de orientadores e a falta de padronização nos metadados sobre os discentes. Essa lacuna pode comprometer a rastreabilidade acadêmica e dificultar análises de redes de pesquisa e colaboração científica. Outro aspecto que não foi discutido na literatura, mas que emerge dos dados, são as oscilações no volume de produção acadêmica ao longo dos anos, sem justificativa evidente nos registros. Outro resultado não abordado na literatura revisada é a variação terminológica na nomenclatura dos trabalhos acadêmicos, dificultando a categorização e a análise de tendências temáticas na área. A literatura aponta a diversificação das linhas de pesquisa, mas não menciona que a falta de um vocabulário padronizado pode comprometer a identificação de padrões. Além disso, os dados sugerem que a pandemia de COVID-19 impactou de maneira expressiva a produção acadêmica, com redução no número de defesas em 2020 e 2021, um impacto que ainda não foi discutido em estudos anteriores sobre a evolução da pós-graduação em Ciência da Informação.

De um modo geral, a análise dos resultados evidencia disparidades na produtividade entre as instituições líderes na área. Embora a literatura reconheça a importância dessas universidades na formação de pesquisadores e no desenvolvimento da área, os dados mostram que algumas instituições mais antigas não necessariamente lideram a produção acadêmica, enquanto outras mais recentes se destacam em volume de trabalhos defendidos. Esse achado sugere que, além do tempo de existência dos programas, fatores como estrutura institucional,

financiamento e redes de colaboração podem ter um papel mais determinante na produtividade acadêmica.

Com base nos resultados obtidos nesta tese, torna-se pertinente compará-los com os estudos anteriores que abordaram a produção científica da Ciência da Informação no Brasil. Embora diversos trabalhos tenham contribuído para a compreensão do campo, muitos se restringiram a recortes específicos, seja em termos temporais, institucionais ou metodológicos. A seguir, apresenta-se um quadro comparativo que destaca as principais características dessas pesquisas, evidenciando os avanços proporcionados por esta tese em relação a escopo, abordagem metodológica, base de dados utilizada e originalidade dos resultados.

Quadro 104 – Comparação entre esta tese e estudos bibliométricos analisados na revisão de literatura (cap. 2) sobre a produção científica em Ciência da Informação no Brasil

Fonte/ Categoria	Silva (2018)	Viana (2016)	Telmo (2018)	Lança, Amaral e Gracioso (2018)	Moreira e Ferneda (2020)	Silva (2021)	Presente tese
Tipo de fonte	Dissertação de Mestrado	Dissertação de Mestrado	Artigo científico	Artigo científico	Artigo científico	Dissertação de Mestrado	Tese de Doutorado
Abrangência institucional	IES do Estado de São Paulo	PPGCI/UFGM e PPGCI/UNESP	PPGCI/UFPB	23 PPGCI brasileiros	20 programas das áreas de informação	Pós-graduação em Ciência da Informação (Brasil)	Todos os programas de pós-graduação em CI no Brasil
Período analisado	2010–2016	Sem recorte temporal explícito (2010–2012 sugerido)	2015–2018	2013–2016	Não explicitado, dados contemporâneos	2010–2020	2004–2022
Base de dados utilizada	Plataforma Lattes, ResearchGate, Google Acadêmico	Repositórios institucionais das IES	Formulários próprios	Currículos Lattes	Currículos Lattes	Scopus e Web of Science	CAPES - Dados abertos (CTD, docentes, discentes, programas)
Recursos computacionais	Não especificado	Não especificado	Excel, Vantage Point, Ucinet, NetDraw	Bibliometria, análise de redes	Não especificado	Bibliometria, VOSviewer	Google Colab, Python, Pandas, bibliotecas auxiliares
Indicadores analisados	Produção científica, redes de colaboração	Temas de teses por categorias temáticas	Participação em bancas, redes de colaboração	Multidisciplinaridade e redes de colaboração	Produtividade científica por programa e canal	Produção científica, coautorias, redes de colaboração	Produção, autoria, distribuição geográfica, perfil docente/discente, temas, banca, tempo de titulação, internacionalização

Fonte/ Categoria	Silva (2018)	Viana (2016)	Telmo (2018)	Lança, Amaral e Gracioso (2018)	Moreira e Ferneda (2020)	Silva (2021)	Presente tese
Recorte temático	Docentes de graduação e pós em Biblioteconomia	Temas das teses em CI	Bancas de doutorado em CI	Interdisciplinaridade na CI	Produção científica em áreas de informação	Produção científica brasileira em Ciência da Informação	Caracterização ampla da CI na pós-graduação brasileira
Análise de qualidade dos dados	Não	Não	Não	Não	Sim (reconhecimento de limitações do Lattes e duplicidade de vínculos)	Não	Diagnóstico de falhas e inconsistências nos dados da CAPES; proposta de aplicação dos princípios FAIR
Proposta metodológica estruturada	Sim (procedimentos definidos, sem proposta metodológica nova)	Não (análise categorial descritiva)	Sim (uso de SNA com ferramentas específicas)	Não (uso bibliométrico clássico, sem proposta nova)	Não (uso de indicadores com base em Lattes)	Sim (procedimentos definidos, sem proposta metodológica nova)	Sim. Formalização da proposta MITAB-DA: Metodologia Integrada de Tratamento e Análise Bibliométrica de Dados Abertos
Originalidade reconhecida	Ênfase em teses, mas com escopo limitado	Diagnóstico inicial do campo	Exploração de relações entre atores	Exploração de relações entre atores	Diagnóstico inicial do campo	Ênfase em teses, mas com escopo limitado	Integração entre dados, perfis institucionais e análise temática com base em dados abertos, com proposta metodológica de curadoria de dados MITAB-DA

Fonte: dados da pesquisa (2025).

Dada essa comparação apresentada no Quadro 104, a originalidade desta pesquisa não reside apenas na adoção de dados abertos da CAPES como fonte primária, mas na forma como esses dados foram sistematicamente organizados, curados e analisados por meio da metodologia MITAB-DA. Foi possível integrar variáveis institucionais, autorais, geográficas e temáticas em um modelo analítico unificado, voltado à compreensão aprofundada da dinâmica da pós-graduação em Ciência da Informação no Brasil. Como resultado, foram identificados padrões e tendências até então pouco explorados, como os efeitos concretos da pandemia de COVID-19 na curva de defesas, oscilações temporais que não se explicam apenas pelo crescimento da área, e casos em que programas recentes superam em produtividade aqueles de longa tradição. Esses achados empíricos, associados à proposta metodológica, conferem à pesquisa um caráter inédito tanto em termos de abordagem quanto de resultados, e contribuem para ampliar a capacidade analítica das avaliações futuras da área.

6.4 Ciência da Informação na Pós-Graduação Brasileira

A análise dos dados da CAPES sobre a pós-graduação no Brasil indica um crescimento contínuo do número de programas, discentes e docentes ao longo das últimas duas décadas. Esse movimento também se reflete na Ciência da Informação, que acompanhou a tendência geral de ampliação. O aumento do número de dissertações e teses defendidas na área demonstra uma consolidação progressiva da CI como campo acadêmico. No entanto, ao comparar com o crescimento de áreas mais tradicionais, como Ciências da Saúde e Engenharias, percebe-se que a expansão da CI ocorre de forma mais moderada, possivelmente em função da menor disponibilidade de programas e do número reduzido de pesquisadores dedicados à área.

A relação entre o crescimento da pós-graduação em geral e o da CI sugere que fatores institucionais, políticas de financiamento e demandas específicas de cada campo exercem influência sobre a velocidade de expansão dos programas. A CAPES classifica os programas de pós-graduação em nove grandes áreas do conhecimento e duzentas e quarenta e oito áreas do conhecimento, o que torna inviável uma comparação detalhada com todas as demais áreas neste trabalho. No entanto, essa análise pode ser aprofundada em estudos futuros, explorando as particularidades do crescimento da CI em relação a outros campos científicos.

A distribuição regional dos programas de pós-graduação em Ciência da Informação apresenta diferenças em relação à distribuição dos programas de outras áreas do conhecimento. Os dados indicam uma concentração expressiva dos programas da área nas

regiões Sudeste e Sul, refletindo um padrão observado na pós-graduação brasileira como um todo. No entanto, em comparação com áreas, como Ciências da Saúde e Engenharias, a presença da CI em determinadas regiões é mais limitada. O número de programas na região Norte, por exemplo, permanece reduzido, enquanto no Nordeste observa-se uma distribuição um pouco mais equilibrada, mas ainda inferior à das regiões mais desenvolvidas economicamente. Esses dados indicam que, apesar do crescimento da pós-graduação na CI, sua distribuição não acompanha de maneira homogênea o avanço geral da pós-graduação brasileira.

Assim, os dados apontam para uma assimetria que, para além de estatística, levanta questões estruturais sobre o modelo atual de avaliação e financiamento da pós-graduação brasileira. Sob uma perspectiva político-institucional, a lógica atual de avaliação da CAPES, centrada em indicadores como produtividade científica, internacionalização e inserção social, tende a beneficiar programas já consolidados, localizados em regiões com maior densidade científica. Isso acaba por reforçar o chamado efeito Mateus, analisado criticamente por Merton (1968), que mostra implicações mais amplas no sistema de comunicação científica. Segundo o autor, os programas mais estruturados recebem mais reconhecimento e recursos, enquanto os emergentes, situados em regiões periféricas, enfrentam maiores dificuldades para atingir os padrões exigidos. Com isso, perpetuam-se a concentração de recursos e talentos científicos.

Sob a ótica da justiça territorial e da política científica, tal concentração compromete o princípio de equidade na produção e acesso ao conhecimento científico. Como argumenta Dagnino (2008), a ciência deve ser compreendida como um ativo estratégico para o desenvolvimento local. O autor defende uma reorientação da ciência brasileira, de modo que ela se articule com as necessidades locais, promova a valorização de recursos e saberes territoriais e fortaleça políticas públicas voltadas ao bem comum. Entende-se que a ausência de programas de Ciência da Informação em certas regiões impede a formação de quadros técnicos especializados, limita a produção de diagnósticos contextualizados e restringe a capacidade dessas localidades de desenvolver soluções informacionais adequadas à sua realidade socioeconômica e cultural.

Portanto, os resultados obtidos nesta pesquisa sugerem repensar os critérios de avaliação e fomento da pós-graduação brasileira, no sentido de incorporar variáveis territoriais, contextuais e estruturais. A superação das assimetrias regionais deve ser tratada como pauta estratégica, tanto para ampliar o acesso à formação científica quanto para fortalecer a pluralidade epistemológica da área. Essa ampliação territorial não é apenas uma

medida de correção, mas uma oportunidade de enriquecer a produção científica nacional com olhares e problemas oriundos de diferentes realidades.

Outra diferença relevante na distribuição dos programas de Ciência da Informação em relação a outras áreas refere-se à concentração institucional. Enquanto programas de outras áreas estão mais distribuídos em universidades federais e estaduais de grande porte, a CI se estrutura predominantemente em instituições com tradição na área de Humanas e Ciências Sociais Aplicadas. Essas diferenças sugerem a necessidade de um olhar mais detalhado sobre os fatores que influenciam a expansão dos programas e sua inserção regional dentro do sistema de pós-graduação brasileiro.

Com base nos resultados, é possível comparar o tempo médio de titulação dos discentes da Ciência da Informação com a média da pós-graduação brasileira. No mestrado acadêmico, os discentes da Ciência da Informação apresentam um tempo médio de titulação semelhante ao da média nacional, com pequenas variações entre os anos analisados. Já no mestrado profissional, observa-se uma diferença maior, com os discentes da Ciência da Informação concluindo o curso em um tempo ligeiramente superior ao da média nacional. Essa variação pode estar associada à estrutura curricular dos programas da área ou a fatores institucionais que impactam a progressão dos discentes.

No doutorado, os dados indicam que o tempo médio de titulação dos discentes da Ciência da Informação é, em geral, maior do que a média dos doutorandos em outras áreas do conhecimento. Além disso, o número de doutorandos titulados anualmente na Ciência da Informação acompanha a tendência da pós-graduação nacional, porém com oscilações mais acentuadas em determinados períodos. Essas diferenças temporais sugerem que a área possui características particulares no processo de formação dos seus discentes, diferenciando-se, em alguns aspectos, do padrão observado no conjunto geral da pós-graduação brasileira.

Com base nos resultados, a comparação entre a idade de conclusão dos discentes na Ciência da Informação e na pós-graduação em geral revela algumas distinções nos diferentes níveis acadêmicos. No mestrado acadêmico, a idade média dos titulados em CI acompanha a tendência nacional, com uma leve variação em relação à média geral da pós-graduação. No mestrado profissional, observa-se que os discentes da Ciência da Informação apresentam, em média, uma idade de conclusão superior à média nacional. Esse padrão pode estar relacionado ao perfil dos ingressantes nessa modalidade, que frequentemente já possuem experiência no mercado de trabalho antes de ingressar no curso. No doutorado, os discentes da Ciência da Informação também tendem a concluir o curso com uma idade ligeiramente superior à média

nacional, o que pode estar associado à trajetória acadêmica e profissional dos pesquisadores na área.

Essa distribuição etária sugere algumas particularidades na trajetória dos discentes da Ciência da Informação em comparação com os da pós-graduação em geral. Enquanto as médias de titulação no mestrado acadêmico se mantêm próximas, a maior idade observada nos concluintes do mestrado profissional e do doutorado em CI pode indicar um percurso mais prolongado ou uma maior inserção de profissionais que ingressam na pós-graduação após uma trajetória consolidada no mercado de trabalho. Esses padrões reforçam a necessidade de um olhar atento às especificidades de cada área do conhecimento dentro do contexto da pós-graduação brasileira.

A análise comparativa entre os docentes da Ciência da Informação e os da pós-graduação em geral revela diferenças em aspectos como nacionalidade, idade no momento da titulação dos discentes e país da última titulação. Em relação à nacionalidade, os dados mostram que a maioria dos docentes na pós-graduação brasileira são brasileiros. Entre os estrangeiros (4,54%), predominam docentes de Portugal, Argentina, Estados Unidos, França, Espanha e Itália. Na Ciência da Informação, essa distribuição se mantém, mas com uma presença ligeiramente maior de estrangeiros (6,79%). Docentes da Espanha aparecem com destaque, seguidos por profissionais de Cuba, Portugal e França, o que pode indicar uma maior colaboração acadêmica entre programas de pós-graduação da área e instituições desses países .

A idade dos docentes no momento da titulação dos discentes também apresenta variações entre a Ciência da Informação e a pós-graduação como um todo. Enquanto na pós-graduação brasileira há uma distribuição relativamente homogênea das idades dos docentes, na CI há uma maior concentração de docentes em faixas etárias específicas. Esse dado sugere que a formação docente na área pode estar associada a trajetórias acadêmicas diferenciadas, possivelmente influenciadas pelo tempo necessário para consolidação dos programas e pela estrutura do campo da Ciência da Informação no Brasil. Além disso, a relação entre a idade dos docentes e os níveis de orientação acadêmica pode impactar a distribuição da carga docente ao longo dos anos .

O país da última titulação dos docentes na Ciência da Informação também apresenta diferenças quando comparado ao panorama geral da pós-graduação brasileira. Enquanto no conjunto nacional a maioria dos docentes obteve sua última titulação no Brasil (84,58%), na CI (86,13%) há uma proporção ligeiramente menor de docentes cuja titulação final ocorreu no exterior. Portugal, Espanha e França são os países que mais aparecem como locais de

formação desses docentes, o que pode refletir a influência histórica e acadêmica dessas nações na estruturação dos programas de pós-graduação em Ciência da Informação no Brasil. Esse dado também pode estar relacionado à formação de redes internacionais de colaboração científica e intercâmbio acadêmico, que impactam tanto a pesquisa quanto a docência na área.

As informações sobre os docentes da Ciência da Informação evidenciam padrões específicos no perfil dos profissionais que atuam na pós-graduação dessa área. A presença de docentes estrangeiros, a variação na idade no momento da titulação dos discentes e a maior frequência de titulação no exterior são características que diferenciam a CI de outras áreas do conhecimento no Brasil. Tais elementos podem influenciar a configuração dos programas, nas parcerias institucionais e na própria dinâmica de formação dos discentes, sugerindo a necessidade de uma análise mais aprofundada sobre como essas diferenças impactam a evolução da área no contexto da pós-graduação nacional.

6.5 Resumo do Capítulo

O capítulo apresenta uma análise detalhada dos achados da pesquisa, destacando padrões, tendências e desafios na pós-graduação em Ciência da Informação no Brasil entre 2004 e 2022. Essa investigação permitiu a identificação de regularidades e variações nos indicadores acadêmicos, proporcionando uma visão sobre a trajetória da área no sistema de pós-graduação nacional. Os resultados são contextualizados a partir da literatura revisada e das políticas institucionais que orientam o desenvolvimento dos programas de mestrado e doutorado. Aspectos como distribuição regional dos programas, variações na produção acadêmica, diferenças na nomenclatura nos trabalhos e inconsistências nos registros acadêmicos são discutidos à luz das normativas da CAPES e das práticas institucionais que moldam a produção do conhecimento na área. A abordagem exploratória adotada possibilitou a identificação de contrastes e disparidades na evolução da pós-graduação em Ciência da Informação, fornecendo subsídios para reflexões sobre a consolidação do campo.

Ao longo da análise, são apontadas discrepâncias nos dados acadêmicos que afetam a confiabilidade e a interpretação dos resultados, incluindo registros incompletos de discentes e docentes, ausência de identificadores únicos e inconsistências na categorização dos programas. A necessidade de aprimoramento contínuo dos dados se torna evidente diante desses desafios, reforçando a importância da adoção de mecanismos que garantam maior padronização e completude das informações registradas. A incorporação dos princípios FAIR surge como

uma estratégia fundamental para fortalecer a confiabilidade e a acessibilidade dos dados, permitindo análises mais precisas e favorecendo o desenvolvimento da pós-graduação no Brasil.

A análise apresentada neste capítulo evidenciou padrões relevantes na produção científica da pós-graduação em Ciência da Informação, ao mesmo tempo em que revelou limitações estruturais nos dados disponíveis. Esses desafios reforçam a importância de uma abordagem metodológica tecnicamente estruturada, capaz de lidar com grandes volumes de dados abertos e promover análises consistentes, reproduzíveis e aderentes às diretrizes da ciência aberta.

Nessa perspectiva, uma das principais contribuições desta pesquisa foi a formulação de uma proposta metodológica própria, desenvolvida ao longo do processo empírico de extração, tratamento e análise dos dados. Trata-se da Metodologia Integrada de Tratamento e Análise Bibliométrica de Dados Abertos (MITAB-DA), apresentada na seção a seguir como um modelo sistematizado e replicável para estudos bibliométricos baseados em dados abertos públicos, que visa assegurar coerência técnica, transparência e alinhamento com os princípios da ciência aberta.

Com base nessa demanda, apresenta-se na próxima seção uma proposta denominada Metodologia Integrada de Tratamento e Análise Bibliométrica de Dados Abertos (MITAB-DA), concebida para sistematizar os procedimentos adotados nesta pesquisa e oferecer um modelo replicável para estudos bibliométricos baseados em informações públicas. A proposta visa assegurar coerência técnica, transparência e alinhamento com os princípios da ciência aberta.

6.6 Proposta de metodologia

A Metodologia Integrada de Tratamento e Análise Bibliométrica de Dados Abertos (MITAB-DA) constitui um dos principais produtos desta tese. Ela resulta da aplicação prática de procedimentos analíticos aos dados da CAPES e busca oferecer uma estrutura metodológica geral, aplicável a diferentes contextos de pesquisa que utilizem bases governamentais abertas.

A MITAB-DA propõe um conjunto estruturado de etapas que vão desde a coleta e limpeza até a análise e visualização dos dados. Ao sistematizar essas práticas, a metodologia visa garantir a transparência, a confiabilidade e a reprodutibilidade dos estudos baseados em dados públicos. Seu foco está na geração de indicadores quantitativos e na identificação de

padrões, tendências e relações institucionais no campo científico e educacional, com potencial de aplicação em diferentes contextos de pesquisa.

Nesta seção, apresenta-se a MITAB-DA como uma estrutura sequencial e interdependente de procedimentos, priorizando a descrição técnica dos procedimentos empregados. Sua implementação prática, com foco nos dados abertos da CAPES, foi aplicada ao estudo da pós-graduação em Ciência da Informação no Brasil e está detalhada em outras seções desta tese, o que ilustra sua aplicabilidade. A seguir, descrevem-se as etapas que compõem a metodologia.

Etapa 1 – Coleta e consolidação de dados abertos

Nesta etapa inicial, realiza-se a identificação e o *download* dos conjuntos de dados disponíveis em plataformas públicas, como o Portal de Dados Abertos da CAPES e repositórios institucionais. A coleta prioriza arquivos em formatos abertos e estruturados, como CSV, por sua compatibilidade com ferramentas computacionais de análise e por favorecerem a reprodutibilidade científica. Durante essa etapa, são selecionadas variáveis fundamentais para a caracterização e análise bibliométrica da produção acadêmica, incluindo informações sobre autoria, titulação, programas, instituições, orientadores, títulos e resumos dos trabalhos, área de avaliação, modalidade do curso e data de defesa. Os metadados desempenham papel central nesse processo, pois viabilizam o reconhecimento, o rastreamento e a contextualização de cada registro, permitindo que os dados possam ser corretamente identificados, classificados e agrupados em fases posteriores. Além disso, essa etapa incorpora critérios legais e éticos para o uso das informações públicas, observando os princípios dos Dados Abertos Governamentais (DAG) e os dispositivos da Lei Geral de Proteção de Dados Pessoais (LGPD). Ao final da coleta, os dados são consolidados em arquivos únicos e organizados em estruturas que permitam seu tratamento sistemático e análise subsequente.

Etapa 2 – Tratamento e normalização dos dados

Concluída a etapa de coleta, inicia-se o processo de tratamento e normalização dos dados, que garante a consistência e a qualidade dos conjuntos utilizados nas análises subsequentes. Essa etapa envolve a aplicação de rotinas computacionais para detecção e remoção de registros duplicados, padronização de formatos e conversão de variáveis para tipos apropriados, especialmente nos casos de datas, que são ajustadas para permitir o cálculo de indicadores como idade e tempo de titulação. Erros de preenchimento, como grafias

divergentes ou variações acidentais em valores categóricos e numéricos, também são tratados de forma sistemática. Para a automação e escalabilidade dessas tarefas, são empregadas linguagens de programação e bibliotecas específicas, com destaque para o uso do Python e bibliotecas como pandas, numpy, unidecode, re (expressões regulares) e datetime. Além disso, o OpenRefine é utilizado como ferramenta auxiliar para inspeção visual, transformação de dados em larga escala e refinamento de padrões inconsistentes. Considerando a ausência de identificadores únicos nos dados originais, essa etapa também contempla a criação de chaves técnicas internas que possibilitam o rastreamento e a vinculação dos registros entre diferentes conjuntos de dados. O objetivo final é garantir que os dados estejam estruturados de maneira uniforme, confiável e pronta para análise.

Etapa 3 – Consolidação e criação de DataFrames estruturados

Após o tratamento e a normalização, os dados são organizados em estruturas analíticas conhecidas como DataFrames, que funcionam como tabelas relacionais e estruturadas, cada uma dedicada a um eixo específico da análise, como programas, docentes, discentes e trabalhos acadêmicos. Essa divisão permite maior clareza na manipulação dos dados, especialmente em projetos que envolvem grandes volumes de registros. A segmentação modular dos dados facilita tanto a realização de análises individuais, por categoria, quanto a integração entre os conjuntos por meio de chaves de relacionamento. Por exemplo:

- *DF1_trabalhos_BR* (teses e dissertações);
- *DF2_programas_BR* (dados dos programas);
- *DF3_docentes_BR* (dados dos docentes);
- *DF4_discentes_BR* (dados dos discentes).

Essa estrutura modular permite análises específicas por categoria e facilita o cruzamento de informações entre os diferentes conjuntos. Cada DataFrame é enriquecido com variáveis derivadas (como tipo de vínculo, região geográfica, área de avaliação da CAPES, entre outros), preparando os dados para as análises quantitativas e relacionais.

Etapa 4 – Análise exploratória e descritiva

Nesta etapa são aplicadas técnicas de análise estatística descritiva para examinar a distribuição dos dados, identificar padrões iniciais e verificar a qualidade dos registros. São utilizados procedimentos como cálculo de frequências absolutas e relativas, medidas de tendência central, séries temporais e histogramas. Além disso, iniciam-se as análises

bibliométricas propriamente ditas, como produtividade por autor, programa e instituição, além de métricas de coautoria e orientações. Essa etapa fornece uma visão geral da estrutura da produção científica e orienta os recortes analíticos posteriores.

Etapa 5 – Recortes temáticos

Com base na estrutura dos dados e nos objetivos da pesquisa, pode ser realizado recorte temático nos conjuntos analisados, como o utilizado neste trabalho, que consistiu na delimitação da área de Ciência da Informação dentro dos dados abertos disponibilizados pela CAPES. Esse tipo de segmentação tem como finalidade isolar subconjuntos específicos da base geral, permitindo uma análise mais direcionada e alinhada ao escopo da investigação. O recorte pode ser operacionalizado por meio de filtros aplicados sobre atributos como área de avaliação, títulos, resumos, palavras-chave, vínculos institucionais, localização geográfica ou perfil dos participantes. Essa etapa viabiliza comparações intra e interinstitucionais, além de possibilitar a identificação de dinâmicas particulares de programas, linhas de pesquisa ou grupos de docentes. Ao estruturar recortes temáticos, torna-se possível compreender padrões específicos de produção científica, mapear tendências e observar especializações locais ou regionais em determinados campos do conhecimento.

Etapa 6 – Visualização dos dados

A visualização de dados constitui um componente central da metodologia MITAB-DA, uma vez que possibilita representar graficamente os resultados obtidos, facilitando a identificação de padrões, contrastes e tendências não evidentes nas análises tabulares. A partir dos dados previamente tratados e organizados, são elaboradas representações visuais como gráficos de linhas, barras, mapas geográficos, diagramas de rede, mapas conceituais, redes de coautoria e colaboração, nuvens de palavras e *heatmaps*. Tais visualizações apoiam a interpretação dos dados ao evidenciar aspectos como lacunas temáticas, assimetrias regionais, mudanças nas dinâmicas institucionais e potenciais estratégias de desenvolvimento da área analisada. Diversas ferramentas e ambientes computacionais podem ser utilizados nessa etapa, com destaque para bibliotecas da linguagem Python (como Matplotlib, Seaborn e Plotly), o *software* IRaMuTeQ (utilizado para análise textual e construção de mapas de similitude), além de plataformas especializadas como VOSviewer e Gephi, voltadas à análise e visualização de redes de coautoria e colaboração científica. Também se destacam plataformas web como Flourish Studio e Datawrapper, que oferecem recursos interativos para a criação de mapas, gráficos e tabelas dinâmicas, ampliando a capacidade de comunicar os resultados de forma clara, acessível e adaptável a diferentes públicos e contextos acadêmicos.

Etapa 7 – Discussão interpretativa dos resultados

Com base nas análises quantitativas e nas visualizações geradas, os resultados são discutidos de forma interpretativa, relacionando-os ao contexto histórico, político e acadêmico do recorte em estudo. São investigadas convergências e divergências entre programas e regiões, bem como tendências emergentes na produção científica. Também se exploram lacunas temáticas, assimetrias de distribuição e oportunidades de colaboração, contribuindo para reflexões sobre políticas públicas, formação de pesquisadores e estratégias de desenvolvimento institucional.

De modo sintetizado, tem-se o quadro-síntese da MITAB-DA, com as principais informações organizadas por etapa, objetivo e procedimentos:

Quadro 105 – Síntese da metodologia MITAB-DA

Etapa	Objetivo	Procedimentos Principais
1. Coleta e consolidação de dados abertos	Reunir dados públicos relevantes para análise bibliométrica	- Identificação e download em portais como o da CAPES - Seleção de variáveis essenciais (autoria, titulação, resumos etc.) - Observância à LGPD e aos princípios dos DAG - Consolidação em arquivos únicos estruturados
2. Tratamento e normalização dos dados	Garantir a qualidade, consistência e padronização dos dados	- Eliminação de duplicatas - Correção de erros e padronização de formatos - Uso de Python (pandas, numpy, re etc.) e OpenRefine - Criação de chaves técnicas internas para vinculação de dados
3. Consolidação e criação de DataFrames estruturados	Estruturar os dados em blocos analíticos temáticos	- Organização dos dados em DataFrames (trabalhos, docentes, programas, discentes) - Inclusão de variáveis derivadas - Facilitação de análises modulares e cruzadas
4. Análise exploratória e descritiva	Examinar padrões gerais e orientar investigações específicas	- Estatísticas descritivas (frequências, médias, histogramas) - Indicadores de produtividade (autores, instituições, programas) - Início das análises bibliométricas
5. Recortes temáticos	Direcionar o foco analítico conforme o escopo da pesquisa	- Filtragem por atributos como área CAPES, palavras-chave e localização - Isolamento de subconjuntos para análises comparativas e especializadas
6. Visualização dos dados	Representar visualmente padrões e relações nos dados	- Gráficos, mapas, diagramas, redes de coautoria e colaboração - Uso de matplotlib, seaborn, plotly, VOSviewer, Gephi, IRaMuTeQ, Flourish, Datawrapper
7. Discussão interpretativa dos resultados	Relacionar achados aos contextos acadêmico e político	- Análise de convergências/divergências regionais e temáticas - Identificação de lacunas, oportunidades e implicações para políticas públicas

Fonte: dados da pesquisa (2025).

A MITAB-DA constitui-se como uma abordagem metodológica replicável para estudos bibliométricos baseados em dados abertos, em especial aqueles disponibilizados pela

CAPES. Sua aplicação favorece a transparência, a reprodutibilidade científica e a geração de conhecimento estratégico sobre a produção acadêmica, especialmente no contexto da pós-graduação brasileira. É uma proposta em consonância com os princípios da ciência aberta e do governo digital, alinhada às demandas contemporâneas por gestão baseada em evidências.

Embora a aplicação empírica da MITAB-DA tenha se dado no domínio da Ciência da Informação, sua estrutura modular e adaptável permite a replicação em outras áreas do conhecimento, desde que sejam utilizadas bases públicas com estrutura semelhante, como os dados da CAPES.

Finalizado este capítulo de discussão dos resultados e apresentação da proposta MITAB-DA, passa-se para o Capítulo Sete, de Conclusões.

7 CONCLUSÕES

A presente pesquisa abordou o mapeamento da produção científica da pós-graduação brasileira no período de 2004 a 2022, com foco específico na área da Ciência da Informação. A partir da análise de dados abertos sobre dissertações e teses disponibilizadas pela CAPES, buscou-se compreender os padrões, tendências e características da produção acadêmica no contexto da pós-graduação. Trata-se de um tema relevante para a área, pois permite observar dinâmicas institucionais, fluxos de conhecimento e transformações nos processos de formação científica, oferecendo subsídios empíricos para a compreensão da evolução da própria área no país.

No âmbito da linha de pesquisa Arquitetura & Organização do Conhecimento (AOC), a investigação dos dados da pós-graduação está alinhada aos princípios que orientam os estudos sobre representação da informação, métricas da produção científica, repositórios digitais e sistemas de organização do conhecimento. Ao trabalhar com um *corpus* de dados institucionais e acadêmicos, esta pesquisa contribui para a consolidação de práticas metodológicas e analíticas voltadas à organização, tratamento, análise e recuperação da informação científica, ao mesmo tempo em que reforça a importância dos estudos quantitativos e qualitativos na compreensão da comunicação científica e de seus suportes organizacionais.

O objetivo geral da pesquisa consistiu em analisar a pós-graduação em Ciência da Informação no Brasil com base em dados abertos de teses e dissertações da CAPES, caracterizando o campo e posicionando-o em relação ao sistema nacional por meio de um estudo bibliométrico, com a aplicação de procedimentos metodológicos de tratamento, organização e análise de dados. As análises expostas no capítulo 5 e as discussões dos resultados apresentadas no capítulo 6 evidenciam que esse objetivo foi alcançado. O estudo possibilitou a visualização do crescimento da área, a identificação de suas principais instituições, a distribuição geográfica dos programas, o perfil de docentes e discentes, bem como o comportamento dos indicadores de titulação. Assim, o estudo cumpriu sua proposta de oferecer uma leitura analítica e crítica da realidade dos dados sobre a pós-graduação em Ciência da Informação, fundamentada em dados concretos e estruturados, ainda que se reconheça a necessidade de aperfeiçoamento contínuo nos sistemas de dados para apoiar futuras investigações.

O primeiro objetivo específico consistiu em identificar os programas brasileiros de pós-graduação na área de Ciência da Informação, com vistas a verificar a quantidade e a

distribuição regional. Esse objetivo foi alcançado com a extração e tratamento dos dados disponibilizados pela CAPES, sendo possível mapear os programas ativos entre 2004 e 2022, classificando-os por modalidade (mestrado acadêmico, mestrado profissional e doutorado acadêmico) e identificando as respectivas instituições responsáveis. A análise evidenciou uma distribuição regional desigual, com maior concentração de programas nas regiões Sudeste e Sul e menor presença nas regiões Norte e Centro-Oeste.

O segundo objetivo específico buscou identificar convergências e divergências entre as regiões brasileiras, com o intuito de verificar tendências e oportunidades de colaboração. Esse objetivo foi parcialmente alcançado. Os dados revelaram desigualdades entre as regiões quanto à quantidade de programas, volume de produção acadêmica e número de titulados, confirmando a predominância do eixo Sudeste-Sul. No entanto, como a pesquisa não investigou diretamente redes de colaboração ou iniciativas interinstitucionais, não foi possível identificar, com base nos dados disponíveis, oportunidades concretas de articulação entre os programas regionais. Ainda assim, a análise evidenciou assimetrias que podem orientar políticas de fomento e expansão, especialmente no que se refere à implementação de programas de mestrado e doutorado em Ciência da Informação nas regiões Norte e Centro-Oeste, onde a presença ainda é reduzida. Esses dados sugerem caminhos possíveis para o fortalecimento do sistema nacional de pós-graduação, com foco na equidade regional.

O terceiro objetivo buscou levantar dados que caracterizassem os PPG em Ciência da Informação, de modo a permitir análises correlacionais entre programas brasileiros. Esse objetivo foi parcialmente alcançado. Diversas variáveis relacionadas aos programas, docentes e discentes — como tempo de titulação, perfil etário, formação acadêmica e volume de produção — foram coletadas, organizadas e analisadas. Embora essas informações tenham possibilitado comparações entre os programas, a análise foi conduzida de forma predominantemente descritiva, sem o emprego de métodos estatísticos que permitissem explorar de modo mais sistemático as correlações entre os conjuntos de dados²⁴.

O quarto objetivo teve como foco distinguir os elementos que caracterizam a Ciência da Informação brasileira, relacionando-os ao contexto mais amplo da pós-graduação nacional. Esse objetivo foi alcançado, pois a comparação entre os dados da área e os dados gerais da pós-graduação permitiu identificar aspectos específicos da Ciência da Informação, como o tempo médio de titulação, o perfil etário predominante dos discentes e a menor incidência de titulações no exterior entre os docentes. Essas singularidades foram discutidas ao longo do

²⁴ A correlação entre os conjuntos de dados não foi possível devido ao fato de não haver chaves que pudessem encontrar os dados de um conjunto em outro conjunto.

capítulo 5, de resultados, e do capítulo 6, de discussões, contribuindo para posicionar a área no panorama da pós-graduação brasileira.

O quinto objetivo específico buscou levantar e analisar dados sobre os programas de pós-graduação, seus docentes e discentes, com o propósito de conhecer seus principais atributos. Esse objetivo foi atingido, uma vez que a partir dos dados da CAPES foi possível traçar o perfil dos atores envolvidos, considerando variáveis como nacionalidade, idade, titulação, formação acadêmica e participação em defesas. A estruturação desses dados nos blocos analíticos permitiu uma caracterização detalhada da pós-graduação em Ciência da Informação entre 2004 e 2022.

Em síntese, os objetivos específicos da pesquisa foram majoritariamente cumpridos. O mapeamento e a distribuição dos programas foram realizados com precisão; as diferenças regionais foram identificadas, embora sem aprofundamento nas possibilidades de cooperação. A caracterização dos programas, docentes e discentes foi conduzida de forma estruturada, ainda que a análise correlacional tenha se mantido em nível descritivo. A comparação com o sistema nacional evidenciou as singularidades da área, e o retrato dos sujeitos envolvidos proporcionou uma visão abrangente do cenário da pós-graduação em Ciência da Informação no período investigado.

A questão de pesquisa — Como mapear e caracterizar a produção científica da pós-graduação em Ciência da Informação no Brasil, com base nos dados abertos da CAPES, e quais procedimentos metodológicos são necessários para garantir a confiabilidade e a replicabilidade dessa análise? De que forma é possível mapear e caracterizar a produção científica dos programas de Pós-Graduação em Ciência da Informação no Brasil, a partir das teses e dissertações defendidas — foi respondida por meio da construção e análise de um conjunto robusto de dados provenientes dos sistemas abertos da CAPES. Para tanto, adotou-se uma proposta metodológica baseada em procedimentos bibliométricos e em etapas sistemáticas de coleta, tratamento, organização e visualização de dados, articuladas com o uso de DataFrames temáticos voltados à área da Ciência da Informação.

A partir da extração de informações sobre programas, discentes, docentes e trabalhos defendidos entre 2004 e 2022, os dados foram progressivamente refinados, padronizados e segmentados, permitindo uma leitura estruturada da pós-graduação na área. As análises realizadas possibilitaram identificar tendências de crescimento, distribuição geográfica, perfis docentes e discentes, além de aspectos específicos da produção intelectual dos programas.

Adicionalmente, foi possível estabelecer comparações com a pós-graduação brasileira como um todo, evidenciando singularidades da Ciência da Informação em relação ao número

de programas, tempo médio de titulação, perfil etário dos discentes, formação dos docentes e volume de produção. Dessa forma, os procedimentos adotados e os dados analisados forneceram uma base empírica e metodológica adequada, capaz de sustentar um panorama confiável e representativo da produção científica da área no período investigado.

Como resultado dessa trajetória empírica e técnica, estruturou-se, na seção 6.6, a MITAB-DA (Metodologia Integrada de Tratamento e Análise Bibliométrica de Dados Abertos), que sistematiza os procedimentos aplicados em um modelo replicável, tecnicamente rigoroso e alinhado à ciência aberta. Essa proposta metodológica emergiu da experiência prática com os dados da CAPES e representa uma contribuição original da pesquisa, com potencial de ser reaplicada em outras áreas do conhecimento e estudos baseados em dados abertos.

As limitações enfrentadas ao longo da pesquisa decorreram, em grande parte, da própria estrutura dos dados abertos disponibilizados pela CAPES. Embora a base ofereça um volume expressivo de informações, foram identificados entraves que afetaram o fluxo das análises. Uma das principais dificuldades foi a ausência ou inconsistência de identificadores únicos para docentes, discentes e instituições, o que dificultou a vinculação precisa entre produções, programas e autores ao longo do tempo. Tal limitação compromete, por exemplo, a rastreabilidade de orientações, a análise de trajetórias acadêmicas e o mapeamento de vínculos entre orientadores e seus orientandos em diferentes períodos.

Também foram identificados campos com preenchimento inadequado, como registros alfanuméricos em colunas que deveriam conter apenas valores numéricos, exigindo um esforço adicional de limpeza e padronização para tornar os dados tratáveis. A nomenclatura dos programas apresentou falta de padronização, sobretudo quando o nome do curso incorporava a área de avaliação — como em “MEDICINA (UROLOGIA)” —, gerando ambiguidades na agregação correta por área. Além disso, observou-se duplicidade de nomes entre programas distintos e diferentes formas de escrita para uma mesma instituição ao longo dos anos. A não obrigatoriedade do preenchimento de certos campos, como palavras-chave ou número de páginas, resultou em registros incompletos. Ademais, o impacto da pandemia de COVID-19 afetou a regularidade da produção acadêmica entre 2020 e 2022, interferindo na comparabilidade temporal dos dados mais recentes.

Reconhece-se, também, como limitação da presente pesquisa, a não realização do cruzamento empírico entre temas das dissertações e teses, linhas de pesquisa e atuação dos orientadores, devido à ausência de metadados padronizados na base da CAPES. No entanto, ao explicitar essa lacuna, já apontada por outros autores (Queiroz; Noronha, 2004; Viana,

2016; Lança; Amaral; Gracioso, 2018), e ainda os desafios técnicos envolvidos, a tese contribui para orientar futuras investigações que aprofundem essa relação fundamental para a avaliação dos programas. Essas limitações evidenciam a importância do aprimoramento contínuo da política de dados abertos da CAPES, especialmente no que diz respeito à qualidade, integridade e interoperabilidade das informações disponibilizadas para fins de pesquisa.

Diante dos resultados alcançados, um primeiro desdobramento possível de estudo futuro diz respeito à análise comparativa entre os próprios programas de pós-graduação em Ciência da Informação ou em outras áreas do conhecimento. Embora esta pesquisa tenha caracterizado a área de forma agregada, estudos futuros podem aprofundar as diferenças entre os programas, explorando títulos, resumos e demais características dos trabalhos defendidos. A aplicação de técnicas de análise de conteúdo ou mineração de texto pode revelar especializações temáticas locais, vocações institucionais e subáreas da CI com maior ou menor grau de desenvolvimento.

Outro tópico promissor de investigação futura está no mapeamento temático da produção científica. A análise mais minuciosa das dissertações e teses pode evidenciar a permanência de temáticas clássicas da área, o surgimento de novas abordagens ou a consolidação de temas emergentes. Esse tipo de estudo pode subsidiar o planejamento estratégico dos programas e orientar políticas de fomento mais sintonizadas com as transformações sociais, tecnológicas e informacionais contemporâneas.

Além disso, o fortalecimento das análises comparativas entre a Ciência da Informação e outras áreas do conhecimento configura-se como relevante. Sugerem-se estudos que adotem recortes por grandes áreas ou por áreas de avaliação que podem identificar padrões metodológicos, variações institucionais e diferenças nos indicadores de titulação, internacionalização e impacto. Tais comparações podem permitir posicionar a Ciência da Informação no conjunto da pós-graduação brasileira, evidenciando suas especificidades e seu grau de articulação com o sistema nacional de ciência, tecnologia e inovação.

Outro estudo futuro potencial é o uso de ferramentas de inteligência artificial (IA) na análise dos dados massivos disponibilizados pela CAPES. A elevada quantidade de registros e a diversidade de variáveis indicam que técnicas de aprendizado de máquina e processamento de linguagem natural podem ser aplicadas para classificar, agrupar e inferir padrões complexos. A adoção dessas tecnologias pode otimizar o processamento dos dados e oferecer novas perspectivas para a compreensão das dinâmicas acadêmicas e científicas da pós-

graduação, tanto na área da Ciência da Informação quanto em outros campos do conhecimento.

A originalidade desta pesquisa reside na articulação entre fundamentos da Ciência da Informação e metodologias da ciência de dados aplicadas à análise da pós-graduação brasileira, com foco específico na área de Ciência da Informação. Ao estruturar um modelo analítico baseado nos dados abertos da CAPES, a tese demonstra a viabilidade do uso de bases públicas para estudos de metaciência em larga escala. Essa abordagem não apenas amplia as possibilidades de compreensão sobre a dinâmica da produção científica nacional, como também oferece uma contribuição metodológica replicável, capaz de ser adaptada a outras áreas do conhecimento ou a outros recortes institucionais, temáticos e temporais.

Com base na análise dos resultados apresentados no Capítulo 5, é possível identificar os usos práticos decorrentes desta pesquisa. A estruturação e o tratamento de um grande volume de dados provenientes da CAPES possibilitaram a construção de um panorama abrangente da pós-graduação em Ciência da Informação no Brasil. Os dados organizados viabilizam consultas por instituição, programa, docente e discente, além de permitirem análises temporais da produção acadêmica. Na prática, esse conjunto de informações pode subsidiar decisões de gestores institucionais e formuladores de políticas públicas, apoiando ações de planejamento, avaliação e redistribuição de recursos com base em evidências empíricas.

As informações sistematizadas também oferecem subsídios para que os próprios programas de pós-graduação reflitam sobre sua trajetória, posicionamento regional e desempenho em relação ao panorama nacional. Além disso, os dados extraídos podem ser reutilizados em pesquisas acadêmicas, análises temáticas ou estudos comparativos, otimizando o tempo e o esforço envolvidos em etapas de coleta e pré-processamento.

O ambiente de dados estruturado abre ainda a possibilidade de desenvolvimento de ferramentas visuais, como painéis interativos e sistemas de consulta dinâmica, ampliando o acesso à informação por parte da comunidade científica, técnica e da sociedade em geral. E a metodologia aplicada nesta pesquisa evidencia a viabilidade do uso de ciência de dados em estudos de metaciência, demonstrando como métodos computacionais podem ampliar a transparência, a reprodutibilidade e a profundidade analítica na avaliação de sistemas científicos e educacionais.

Como contribuições, esta pesquisa, ao conduzir uma investigação aprofundada da pós-graduação em Ciência da Informação no Brasil no período de 2004 e 2022, oferece subsídios concretos para a compreensão das dinâmicas institucionais, acadêmicas e científicas que

configuram o campo. A sistematização de dados provenientes de fontes públicas possibilitou o mapeamento de aspectos estruturais da área, como a distribuição regional dos programas, a evolução no número de titulados, os perfis de discentes e docentes, além dos padrões temáticos predominantes nos trabalhos defendidos. Essa abordagem revelou configurações específicas da área que podem orientar políticas públicas, práticas de gestão institucional e ações de fortalecimento da pós-graduação.

Do ponto de vista teórico-metodológico, a pesquisa contribui para o avanço dos estudos em ciência aberta e gestão da informação acadêmica, ao demonstrar o potencial analítico dos dados abertos institucionais na avaliação de sistemas complexos como o da pós-graduação. A adoção de procedimentos de tratamento, padronização e integração de grandes volumes de dados, aliada ao uso de estratégias de visualização e análise quantitativa e textual, evidenciou a relevância de métodos computacionais aplicados à Ciência da Informação. Tais aportes metodológicos podem ser replicados ou adaptados em outras investigações voltadas ao monitoramento da produção científica e à avaliação de políticas públicas no campo da educação superior e da ciência e tecnologia.

No que se refere a contribuições para profissionais, os resultados alcançados oferecem uma base para gestores, formuladores de políticas e lideranças acadêmicas, ao evidenciar desigualdades regionais, diferenças de desempenho entre programas e perfis variados de atuação docente. As informações organizadas ao longo da pesquisa podem apoiar o planejamento institucional, a definição de estratégias de expansão e a consolidação de práticas mais equitativas e inclusivas na pós-graduação. Especificamente para os profissionais da informação, o estudo também sinaliza caminhos promissores para o desenvolvimento de instrumentos de organização, indexação e análise de dados científicos, com foco na melhoria da interoperabilidade e da reutilização das informações acadêmicas.

Esta investigação contribui diretamente para a consolidação da Ciência da Informação como campo estratégico no contexto da pós-graduação brasileira. Ao oferecer uma visão ampliada e detalhada de sua produção científica, estrutura e inserção institucional, a pesquisa reforça o papel da área na mediação da informação e na análise de sistemas de conhecimento. As lacunas e desafios identificados — como a ausência de identificadores únicos, a padronização insuficiente dos dados e a baixa articulação entre programas — são elementos trabalhados pela área e configuram oportunidades para novos estudos e intervenções. Assim, esta pesquisa se estabelece como ponto de partida para reflexões futuras, ao mesmo tempo em que reafirma a relevância da Ciência da Informação para o aprimoramento da produção acadêmica nacional.

Do ponto de vista pessoal, esta pesquisa representou a convergência entre uma trajetória consolidada na área da Computação e a busca por novos referenciais teóricos e metodológicos no campo da Ciência da Informação. Ao aplicar conhecimentos técnicos em programação, bancos de dados e análise de dados à realidade complexa da pós-graduação brasileira, o autor pôde ressignificar sua prática profissional, ampliando-a para além das fronteiras da tecnologia. A experiência foi ainda enriquecida pela participação no Programa de Doutorado Sanduíche no Exterior (PDSE), realizado na Áustria, que proporcionou contato com abordagens internacionais de análise de sistemas de informação e políticas públicas em ciência e tecnologia. Esse intercâmbio ampliou o repertório metodológico do autor e reforçou sua perspectiva crítica sobre os desafios da interoperabilidade, da padronização e do uso ético de dados acadêmicos. Assim, a pesquisa constituiu um processo formativo que reafirmou o compromisso com a transformação de dados em informação útil, acessível e socialmente relevante — princípio que motivou e orientou todo o percurso investigativo.

Conclui-se, assim, um percurso que alia técnica e crítica, oferecendo não apenas dados e visualizações, mas também um convite à reflexão sobre como se produz, registra, armazena, acessa, usa e avalia o conhecimento científico no Brasil. Ao integrar saberes da Computação e da Ciência da Informação, esta pesquisa reafirma o valor da informação como base para uma ciência mais transparente, útil e socialmente comprometida.

REFERÊNCIAS

AFONSO, Adriano *et al.* **Folha de Cálculo Calc**: Coleccção Manual TIC e LibreOffice. ISCTE–Instituto Universitário de Lisboa, 2014.

AGRA, Maria Aparecida Costa *et al.* Dissertações e teses da enfermagem acerca do serviço de atendimento móvel de urgência: estudo bibliométrico. **Texto & Contexto-Enfermagem**, v. 27, n. 1, p. e3500016, 2018.

ALMEIDA, Karla Nazareth Corrêa de. **A pós-graduação no Brasil**: história de uma tradição inventada. 2017. 213f. Tese (Doutorado em Educação) – Faculdade de Educação, Universidade Estadual de Campinas, Campinas, 2017. Disponível em: <https://encr.pw/XUZle>. Acesso em: 20 abr. 2022.

ALVAREZ, Gonzalo Rubén; CAREGNATO, Sônia Elisa. A Ciência da Informação e sua contribuição para a avaliação do conhecimento científico. **Biblos**: Revista do Instituto de Ciências Humanas e da Informação, Rio Grande, v. 31, n. 1, p. 9-26, jan./jun. 2017. Disponível em: <https://encr.pw/oGRvT>. Acesso em: 24 abr. 2022.

ALVES, Bruno Henrique; DALESSANDRO, Rafael Cacciolari; SANTOS, Fernanda Bochi dos. Colaboração científica no periódico Knowledge Organization: elementos para caracterização de um domínio. **Knowledge Organization**, [S.l.], v. 6, n. 1, p. 137-144, jan./jun. 2019. Disponível em: <https://11nq.com/tVoGM>. Acesso em: 22 abr. 2022.

ALVES, Francisco Regis Vieira; VIEIRA, Renata Passos Machado. The Newton fractal's Leonardo sequence study with the Google Colab. **International Electronic Journal of Mathematics Education**, [S. l.], v. 15, n. 2, p. 1-9, jul./dez. 2019. Disponível em: <https://encr.pw/e6giF>. Acesso em: 15 out. 2022.

AMARAL, Prioste Eloni *et al.* Uso da análise bibliométrica nos Anais do ENEGEP de 2008 a 2016. **Exacta**, [S. l.], v. 17, n. 4, p. 273-282. out./dez. 2019. Disponível em: <https://11nq.com/V10rZ>. Acesso em: 30 jul. 2022.

ANJOS, Francisco Antônio dos; RODRIGUES, Gilson de Jesus Mota. Bolsa CNPq produtividade em pesquisa: perfil dos pesquisadores da área de Turismo. **Rosa dos Ventos**, Fortaleza, v. 11, n. 1, p. 194-210, jan./jun. 2019. Disponível em: <https://11nq.com/Jx02M>. Acesso em: 20 abr. 2022.

ARAÚJO, Ronaldo Ferreira; ALVARENGA, Lídia. A bibliometria na pesquisa científica da pós-graduação brasileira de 1987 a 2007. **Encontros Bibli**: Revista Eletrônica de Biblioteconomia e Ciência da Informação, Florianópolis, v. 16, n. 31, p.51-70, jan./jun. 2011. Disponível em: <https://11nq.com/f8LZl>. Acesso em: 19 jun. 2022.

ARAÚJO, Carlos Alberto Ávila; VALENTIM, Marta-Ligia Pomim. A Ciência da Informação no Brasil: mapeamento da pesquisa e cenário institucional. **Bibliotecas. Anales de Investigación**, v. 15, n. 2, p. 232-259, 2019.

ARAÚJO, Paula Carina de; MIGUEL, Sandra. Motivações dos discentes do Programa de Pós-Graduação em Direito da Universidade Federal do Paraná (UFPR) para publicar em periódicos científicos no domínio do Direito. **Perspectivas em Ciência da Informação**, Belo

Horizonte, v. 22, n. 1, p. 38-56, jan./jun. 2017. Disponível em: <https://encr.pw/snfGr>. Acesso em: 26 maio 2022.

ARMBRUST, Michael *et al.* Spark sql: Relational data processing in spark. In: ACM SIGMOD INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA, 2015, Melbourne Victoria Australia. **Proceedings...**, New York: Association for Computing Machinery, 2015. p. 1383-1394. DOI: <https://doi.org/10.1145/2723372.2742797>

Arnold, Lisa., Breitmayer, Marius., Reichert, Mandred. Monitoring Object-Centric Business Processes: An Empirical Study. In: Nurcan, S., Opdahl, A.L., Mouratidis, H., Tsohou, A. Research Challenges in Information Science: Information Science and the Connected World. RCIS 2023. **Lecture Notes in Business Information Processing**, vol 476. Springer, Cham. DOI: https://doi.org/10.1007/978-3-031-33080-3_20

ASSOCIAÇÃO NACIONAL DE PESQUISA E PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO (ANCIB). **Grupos de Trabalho**. [S. l.], 2022. Disponível em: <https://11nq.com/30EaJ/>. Acesso em: 24 abr. 2022.

ASSUNÇÃO-LUIZ, Alan Vinícius *et al.* Impacto da Covid-19 em alunos de pós-graduação. **Olhares & trilhas**, v. 23, n. 2, p. 538-554, 2021.

ATTARD, Judie *et al.* A systematic review of open government data initiatives. **Government information quarterly**, v. 32, n. 4, p. 399-418, 2015.

AUDY, Jorge Luís Nicolas(org.). **Evolução do snpg no decênio do pnpg 2011-2020**. Brasília-Df: Capes, 2021. 208 p. Disponível em: <https://acesse.one/WVDDw>. Acesso em: 13 maio 2021.

AUTRAN, Marynice Medeiros Matos *et al.* Perfil de produção acadêmica dos programas brasileiros de pós-graduação em Ciência da Informação 2008-2012. **Perspectivas em Ciência da Informação**, v. 20, n. 4, p. 57-78, 2015.

BACKES, Vânia Marli Schubert *et al.* Teses e dissertações de enfermeiros sobre educação em enfermagem e saúde: um estudo bibliométrico. **Revista Brasileira de Enfermagem**, v. 66, p. 251-256, 2013.

BAGGIO, Maria Aparecida *et al.* Produções de teses e dissertações da enfermagem em Portugal de 2000-2010: estudo bibliométrico. **Texto & Contexto-Enfermagem**, v. 23, p. 250-260, 2014.

BARBALHO, Célia Regina Simonetti *et al.* Curso de bacharelado em biblioteconomia na modalidade a distância: relato da experiência da Universidade Aberta do Brasil/Capes. **Revista Eletrônica da ABDF**, São Paulo, v. 5, n. 1, p. 68-104, jan./jun. 2021. Disponível em: <https://brapci.inf.br/index.php/res/download/166266>. Acesso em: 24 abr. 2022.

BASE DE DADOS REFERENCIAIS DE ARTIGOS DE PERIÓDICOS EM CIÊNCIA DA INFORMAÇÃO (BRAPCI). **Sobre a base**. Curitiba, 2022. Disponível em: <https://brapci.inf.br/index.php/res/about>. Acesso em: 18 jun. 2022.

BATES, Marcia J.; MAACK, Mary Niles. **Encyclopedia of library and information sciences**. CRC Press, Inc., 2009.

BARBOSA, Gabroela da Rocha. **Os Planos Nacionais de Pós-Graduação (PNPG): uma discussão sobre a política de C&T nacional e a formação da agenda de pesquisa**. Brasília: Capes, 2020.

BARBOSA, Maria Lúcia; SEVERO, Carlos Emilio Padilla; REATEGUI, Eliseo. Mineração de padrões no gênero textual blog. **Novas Tecnologias na Educação**, Porto Alegre, v. 7, n. 3, p. 1-11, dez. 2009. Disponível em: <https://11nq.com/H0TMR>. Acesso em: 26 abr. 2022.

BELLUZZO, Regina Célia Baptista. **A competência em informação no Brasil: cenários e espectros**. São Paulo: ABECIN Editora, 2018. 217p.

BEIRA, Joana Carlos *et al.* Indicadores bibliométricos na produção científica em periódicos brasileiros da ciência da informação no estrato A1. **Revista ACB: Biblioteconomia em Santa Catarina**, Florianópolis, v. 25, n. 2, p. 383-408, abr./jul. 2020. Disponível em: <https://11nq.com/8k2fC>. Acesso em: 24 abr. 2022.

BESERRA, Patrícia Josefa Fernandes *et al.* Scientific production of the International Classification for Nursing Practice: a bibliometric study. **Revista Brasileira de Enfermagem**, v. 71, p. 2860-2868, 2018.

BERNERS-LEE, T. **5 [stars] open data**. [S.l.], 2012. Disponível em: <http://5stardata.info/en/>. Acesso em: 10 jan. 2023.

BISWAS, Kusan; CHAND, Satish. Python programming for machine learning and big data analysis: Issues, solutions and benchmarks. In: **Communication and Computing Systems**. [s.l.]: CRC Press, 2019. p. 382-387. [e-book]

BOYD, Danah; CRAWFORD, Kate. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. **Information, communication & society**, v. 15, n. 5, p. 662-679, 2012.

BORBA, Vildeane; CAREGNATO, Sônia. Indicadores alométricos nos periódicos brasileiros em Ciência da Informação: um panorama de pesquisa. **Em Questão**, Porto Alegre, v. 27, n. 3, p. 275-302, jul./set. 2021. Disponível em: <https://11nq.com/VEnYT>. Acesso em: 24 abr. 2022.

BORGMAN, Christine L. **Big data, little data, no data: Scholarship in the networked world**. [s.l.]: MIT press, 2015. ISBN: 9780262529914

BOTELHO, Louise Lira Roedel; CUNHA, Cristiano Castro de Almeida; MACEDO, Marcelo. O método da revisão integrativa nos estudos organizacionais. **Gestão e sociedade**, v. 5, n. 11, p. 121-136, 2011.

BOTELHO, Rafael Guimarães; DE OLIVEIRA, Cristina da Cruz. Literaturas branca e cinzenta: uma revisão conceitual. **Ciência da Informação**, v. 44, n. 3, 2015.

BORKO, Harold. Information Science: what is it? **American Documentation**, [S. l.], v. 19, n. 1, p. 3- 5, jan. 1968.

BRASIL. Decreto n. 29.741, de 11 de julho de 1951. Institui uma Comissão para promover a Campanha Nacional de Aperfeiçoamento de Pessoal de Nível Superior. **Diário Oficial da União**, Brasília, 13 jul. 1951. Disponível em: <https://encurtador.com.br/5y3pz>. Acesso em: 20 abr. 2022.

BRASIL. Lei n. 9.394, de 20 de dezembro de 1996. Estabelece as diretrizes e bases da educação nacional. Brasília, 1996. **Diário Oficial da União**, Brasília-DF, 23 dez. 1996. Disponível em: <https://11nk.dev/uUXhK>. Acesso em: 20 abr. 2022.

BRASIL. Ministério da Educação. Resolução CNE/CES Nº 24, de 18 de dezembro de 2002. Altera a redação do parágrafo 4º do artigo 1º e o artigo 2º, da Resolução CNE/CES 1/2001, que estabelece normas para o funcionamento de cursos de pós-graduação. Brasília, 2002. **Diário Oficial da União**, Brasília-DF, 20 dez. 2002. Disponível em: <https://encurtador.com.br/54wAl>. Acesso em: 30 abr. 2022.

BRASIL, Ministério da Educação. Resolução n. 1, de 8 de junho de 2007. Estabelece normas para o funcionamento de cursos de Pós-Graduação lato sensu, em nível de especialização. **Diário Oficial da União**, Brasília-DF, 2007. Disponível em: https://portal.mec.gov.br/cne/arquivos/pdf/rces001_07.pdf. Acesso em: 20 abr. 2022.

BRASIL. Decreto nº 8.777, de 11 de maio de 2016. Institui a Política de Dados Abertos do Poder Executivo Federal. **Diário Oficial da União**, Brasília-DF, 12 maio 2016. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2016/decreto/d8777.htm. Acesso em: 4 março 2025.

BRASIL. Lei nº 13.415, de 16 de fevereiro de 2017. Altera a Lei nº 9394/96 que estabelece as diretrizes e bases da educação nacional e dá outras providências. **Diário Oficial da União**, Poder Executivo, Brasília-DF, 17 fev. 2017a.

BRASIL. Ministério da Educação. Portaria nº 389, de 23 de março de 2017. Dispõe sobre o mestrado e o doutorado profissionais no âmbito da pós-graduação stricto sensu. **Diário Oficial da União**, seção 1, Brasília, DF, n. 57, p. 21, 24 mar. 2017b. Disponível em: <https://cad.capes.gov.br/ato-administrativo-detallar?idAtoAdmElastic=241>. Acesso em: 27 março 2025.

BRASIL. Coordenação de Aperfeiçoamento de Pessoal de Nível Superior. Portaria nº 256, de 23 de novembro de 2018. Dispõe sobre fusão, desmembramento e migração dos programas de pós-graduação stricto sensu avaliados pela CAPES. **Diário Oficial da União**, Brasília-DF, 28 nov. 2018. Disponível em: <https://cad.capes.gov.br/ato-administrativo-detallar?idAtoAdmElastic=63>. Acesso em: 30 março 2025.

BRASIL. Ministério da Educação. Coordenação de Aperfeiçoamento de Pessoal de Nível Superior. Portaria nº 60, de 20 de março de 2019. Dispõe sobre o mestrado e doutorado profissionais no âmbito da pós-graduação stricto sensu. **Diário Oficial da União**: seção 1, Brasília, DF, n. 56, p. 59, 22 mar. 2019. Disponível em: <https://cad.capes.gov.br/ato-administrativo-detallar?idAtoAdmElastic=884>. Acesso em: 27 março 2025.

BRASIL. **Plano Nacional de Pós-Graduação – PNPG 2024–2028**. Brasília, DF: CAPES, 2023. Disponível em: <https://acesse.one/tlg1G>. Acesso em: 28 mar. 2025

BRILHANTE, Aline Veras Moraes *et al.* Um estudo bibliométrico sobre a violência de gênero. **Saúde e Sociedade**, São Paulo, v. 25, n. 3, p. 703-715, mar. 2016. Disponível em: <https://encr.pw/B3mrG>. Acesso em: 12 out. 2022.

BRITO, Kellyton dos Santos *et al.* Is Brazilian open government data actually open data?: An analysis of the current scenario. **International Journal of E-Planning Research (IJEPR)**, v. 4, n. 2, p. 57-73, 2015. Disponível em: <https://11nq.com/EuKfM>

BRUGGER, Jérôme *et al.* Current barriers to open government data use and visualization by political intermediaries. In: 2016 CONFERENCE FOR E-DEMOCRACY AND OPEN GOVERNMENT (CeDEM). Krems, Áustria: IEEE, 2016. p. 219-229. DOI: 10.1109/CeDEM.2016.18

BUCKLAND, Michael. What kind of science can information science be?. **Journal of the American society for information science and technology**, v. 63, n. 1, p. 1-7, 2012. Disponível em: <https://encr.pw/LsuHX>. Acesso em 12 abr. 2025

BUDLER, Marko; ŽUPIČ, Ivan; TRKMAN, Peter. The development of business model research: A bibliometric review. **Journal of Business Research**, v. 135, p. 480-495, 2021.

BUFREM, Leilah Santiago *et al.* Produção científica em Ciência da Informação: análise temática em artigos de revistas brasileiras. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 12, n. 1, p. 38-49, jan./jun. 2007. Disponível em: <https://11nq.com/2yACk>

CAIRO, Alberto. **The Truthful Art**: data, charts and maps for communication. New Riders: San Francisco, 2016.

CAJAZEIRA, Paulo Eduardo Silva Lins; SILVA, Hernandes Andrade. As publicações em coautoria e colaboração científica em Comunicação na Universidade Federal do Piauí. **Ciência da Informação**, Brasília, DF, v. 50, n. 1, p. 143-160, jan./abr. 2021. Disponível em: <http://revista.ibict.br/ciinf/article/view/5209/5423>. Acesso em: 20 abr. 2022.

CALDAS, Miguel. P.; TINOCO, Tatiana. Pesquisa em gestão de recursos humanos nos anos de 1990: um estudo bibliométrico. **Revista de Administração de Empresas**, on-line, v. 44, n. 3, p. 100-114, 2004.

CAMPELLO, Bernadete Santos; CENDÓN, Beatriz Valadares; KREMER, Jeannette Marguerite. **Fontes de informação para pesquisadores e profissionais**. Belo Horizonte: Editora Ufmg, 2000.

COORDENAÇÃO DE APERFEIÇOAMENTO DE PESSOAL DE NÍVEL SUPERIOR - CAPES. **Plano de dados Abertos**. Brasília: CAPES, 2017. Disponível em: <https://acesse.one/uaO26>. Acesso em: 03 de abr. de 2023

COORDENAÇÃO DE APERFEIÇOAMENTO DE PESSOAL DE NÍVEL SUPERIOR - CAPES.r. **Dados Abertos da CAPES**. Brasília: CAPES, 2022a. Disponível <https://dadosabertos.capes.gov.br/>. Acesso em: 03 de jun. de 2022

COORDENAÇÃO DE APERFEIÇOAMENTO DE PESSOAL DE NÍVEL SUPERIOR - CAPES. **Perguntas e respostas sob aspectos da LGPD**. Brasília: CAPES, 2022b. Disponível em: <https://11nk.dev/mMOTI>. Acesso em: 03 de out. de 2022.

COORDENAÇÃO DE APERFEIÇOAMENTO DE PESSOAL DE NÍVEL SUPERIOR - CAPES. **Avaliação da Pós-graduação**. Brasília: CAPES, 2022c. Disponível em: <https://acesse.one/HIkYk>. Acesso em: 03 de nov. de 2022.

COORDENAÇÃO DE APERFEIÇOAMENTO DE PESSOAL DE NÍVEL SUPERIOR - CAPES. **Plataforma Sucupira**. Brasília: CAPES, 2022d. Disponível em <https://sucupira.capes.gov.br/>. Acesso em: 03 de mar. de 2022.

COORDENAÇÃO DE APERFEIÇOAMENTO DE PESSOAL DE NÍVEL SUPERIOR - CAPES. **Áreas de avaliação**. Brasília: CAPES, 2024. Disponível em: <https://acesse.one/b6Hyc>. Acesso em: 3 de mar. de 2025.

CAPURRO, Rafael. Epistemologia e Ciência da Informação. In: V ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 5., 10 a 14 de novembro de 2003, UFMG, Belo Horizonte, MG. **Anais [...]** Belo Horizonte: UFMG, 2003.

CAPURRO, Rafael. Epistemología y ciencia de la información. **Enlace**, v. 4, n. 1, p. 11-29, 2007. Disponível em: <https://encr.pw/R6L2s>. Acesso em: 12 abr. 2025.

CARELLI, Ana Esmeralda. Coleção de dissertações e teses: contribuição das bibliotecas universitárias na construção do conhecimento científico. **Repositório - FEBAB**, São Paulo, v. 13, 2004. Disponível em: <https://encr.pw/bnhxE>. Acesso em: 27 maio 2023.

CARIBÉ, Rita de Cássia do Vale. Comunicação científica: reflexões sobre o conceito. **Informação & Sociedade: estudos**, João Pessoa, v. 25, n. 3, p. 89-104, set./dez. 2015. Disponível em: <https://11nq.com/3yM3U>. Acesso em: 21 abr. 2022.

CARNEIRO, Felipe Ferreira Barros; FERREIRA NETO, Amarílio; SANTOS, Wagner dos. Redes de colaboração científica em Educação Física: comparação entre a subárea Biodinâmica do Movimento e a subárea Sociocultural e Pedagógica. **Em Questão**, Porto Alegre, v. 26, n. 3, p. 145-164, set./dez. 2020. Disponível em: <https://seer.ufrgs.br/index.php/EmQuestao/article/view/99022.pdf>. Acesso em: 20 abr. 2022.

CARVALHO, Jonathas. Os mestrados profissionais na Ciência da Informação no Brasil. **Revista Biblioo**, Rio de Janeiro, v. 81, n. 4, s.p., abr. 2021. Disponível em: <https://biblioo.info/trajetorias-e-perspectivas>. Acesso em: 27 maio 2022.

CASSAVIA, Nunziato *et al.* Data preparation for tourist data big data warehousing. In: **Special Session on Knowledge Discovery meets Information Systems: Experiences and Lessons Learned Dealing with Real-life Scenarios**. SCITEPRESS, 2014. p. 419-426. DOI: 10.5220/0005144004190426

CASTELLS, Manuel. **A sociedade em rede**. São Paulo: Paz e Terra, 1999.

ÇELIK, Şenol. Bibliometrics Analysis And An Husbandry Practice. **Journal of Multidisciplinary Engineering Science Studies**, [S. l.], v. 6, n. 12, p. 3632-3641, dez. 2020. Disponível em: <https://encr.pw/Flxjq>. Acesso em: 24 abr. 2022.

CENTRO DE GESTÃO E ESTUDOS ESTRATÉGICOS. **Brasil: Mestres e Doutores 2024**. Brasília, DF: CGEE, 2024. Disponível em: <https://mestresdoutores2024.cgee.org.br>

CHEN, Daniel Y. **Análise de dados com Python e Pandas**. São Paulo: Novatec Editora, 2018.

CHEN, Hsinchun; CHIANG, Roger HL; STOREY, Veda C. Business intelligence and analytics: From big data to big impact. **MIS quarterly**, p. 1165-1188, 2012.

CHOO, Chun Wei. **The knowing organization**: how organizations use information to construct meaning, create knowledge and make decisions. New York: Oxford University Press, 1998. Disponível em: <https://encr.pw/fH0F2>. Acesso em: 12 abr. 2025.

CHUEKE, Gabriel Vouga; AMATUCCI, Marcos. O que é bibliometria? Uma introdução ao Fórum. **Revista Eletrônica de Negócios Internacionais**, São Paulo, v. 10, n. 2, p. 1-5, maio/ago. 2015. Disponível em: <https://11nq.com/XII98>. Acesso em: 24 abr. 2022.

COELHO, Ana. **As redes de colaboração internacional de autores**: o impacto na construção do conhecimento e nas revistas do domínio da ciência da informação. 2021. 10f. Artigo (Trabalho de Conclusão de Curso) – Faculdade de Artes e Humanidades, Universidade de Coimbra, Lisboa, 2021. Disponível em: <https://encr.pw/AHSId>. Acesso em: 25 abr. 2022.

COELHO, Flávio Codeço; BRANDÃO, Adeilton. Decentralising scientific publishing: can the blockchain improve science communication? **Memorial Institut Oswaldo Cruz**, Rio de Janeiro, v. 114, n. 1, p. 1-4, jan. 2019. Disponível em: <https://11nq.com/6pYMs>. Acesso em: 20 abr. 2022.

COOKE, Steven *et al.* Considerations for effective scientific communication. **Facets**, [S. l.], v. 2, n. 1, p. 233-248. 2017. Disponível em: <https://11nq.com/IT6oY>. Acesso em: 20 abr. 2022.

COSTA, Teresa; LOPES, Sílvia; FERNÁNDEZ-LLIMÓS, Fernando. A Bibliometria e a Avaliação da Produção Científica: indicadores e ferramentas. In: **Actas do Congresso Nacional de Bibliotecários, Arquivistas e Documentalistas**. Associação Portuguesa de Bibliotecários, Arquivistas e Documentalistas, 2012.

COSTA, Odaléia Alves da; RODRIGUES, Anny Camila Lima. Mapeamento da produção científica na BDTD do IBICT sobre a Pedagogia da Alternância de 2011 a 2018. **Revista Brasileira de Educação do Campo**, v. 4, p. e7257-e7257, 2019.

COSTA, Solange Fátima Geraldo da *et al.* The use of the Calgary model in nursing dissertations and theses: a bibliometric study/Utilização do modelo Calgary em dissertações e teses de enfermagem: estudo bibliométrico. **Revista de Pesquisa Cuidado é Fundamental Online**, v. 10, n. 4, p. 1156-1163, 2018.

CORSAR, David; EDWARDS, Peter. Challenges of open data quality: More than just license, format, and customer support. **Journal of Data and Information Quality (JDIQ)**, v. 9, n. 1, p. 1-4, 2017. Acesso em: 28 mar. 2023.

CUNHA, Ísis Trindade da Silva. **O termo cultura na pesquisa em Ciência da Informação no Brasil**: um estudo da produção científica. 2021. 97f. Dissertação (Mestrado em Ciência da Informação) – Centro de Artes e Comunicação, Universidade Federal de Pernambuco, Recife, 2021. Disponível em: <https://l1nq.com/gFAmI>. Acesso em: 26 abr. 2022.

CRESWELL, John. **Investigação qualitativa e projeto de pesquisa**: escolhendo entre cinco abordagens. Porto Alegre: Penso, 2014.

CRISTÓVAM, José Sérgio da Silva; HAHN, Tatiana Meinhart. Administração Pública orientada por dados: Governo aberto e infraestrutura nacional de dados abertos. **Revista de Direito Administrativo e Gestão Pública**, v. 6, n. 1, p. 1-24, 2020. Acesso em: 05 fev. 2022.

CYR, Sylvio; WEI CHOO, Chun. The individual and social dynamics of knowledge sharing: an exploratory study. **Journal of documentation**, v. 66, n. 6, p. 824-846, 2010. Disponível em: <https://encr.pw/x7aKB>. Acesso em: 12 abr. 2025.

DAGNINO, Renato. As trajetórias dos Estudos sobre Ciência, Tecnologia e Sociedade e da Política Científica e Tecnológica na Ibero-América. **Alexandria: Revista de Educação em Ciência e Tecnologia**, [S. l.], v. 1, n. 2, p. 3–36, jul. 2008. Disponível em: <https://encr.pw/Ui5cc>. Acesso em: 12 abr. 2025.

DEEPA, Natarajan *et al.* A survey on blockchain for big data: approaches, opportunities, and future directions. **Future Generation Computer Systems**, 2022. Disponível em: <https://doi.org/10.1016/j.future.2022.01.017>. Acesso em: 29 mar. 2023.

DENG, Shengli; XIA, Sudi. Mapping of interdisciplinarity in information behavior research: a quantitative study using diversity measurement and co-occurrence analysis. **Scientometrics**, [S. l.], v. 124, n. 1, p. 489-513, jan./jun. 2020. Disponível em: <https://encr.pw/8Wbrt>. Acesso em: 25 abr. 2022.

DI VAIO, Assunta; HASSAN, Rohail; ALAVOINE, Claude. Data intelligence and analytics: A bibliometric analysis of human–Artificial intelligence in public sector decision-making effectiveness. **Technological Forecasting and Social Change**, v. 174, p. 121201, 2022.

DIAS, Kalina Coeli Costa de Oliveira *et al.* Dissertações e teses sobre cuidados paliativos em oncologia pediátrica: estudo bibliométrico. **Acta Paulista de Enfermagem**, v. 33, p. eAPE20190264, 2020.

DIAS, Thiago Magela; MOITA, Gray Farias; DIAS, Patricia Mascarenhas. Um estudo sobre a rede de colaboração científica dos pesquisadores brasileiros com currículos cadastrados na Plataforma Lattes. **Em Questão**, Porto Alegre, v. 25, n. 1, p. 63-86, jan./abr. 2019. Disponível em: <https://encr.pw/5GVhZ>. Acesso em: 20 abr. 2022

DIEHL, Astor Antônio; TATIM, Denise Carvalho. **Pesquisa em ciências sociais aplicadas: métodos e técnicas**. São Paulo: Pearson Prentice Hall, 2004.

DONTHU, Naveen *et al.* Forty years of the International Journal of Information Management: a bibliometric analysis. **International Journal of Information Management**, [S. l.], v. 57, n. 4, p. 1-13, abr. 2021. Disponível em: <https://encr.pw/wXXSX>. Acesso em: 25 abr. 2022.

DROESCHER, Fernanda Dias; SILVA, Edna Lucia da. O pesquisador e a produção científica. **Perspectivas em ciência da informação**, v. 19, p. 170-189, 2014.

EMMANUEL, Tlanelo *et al.* A survey on missing data in machine learning. **Journal of Big Data**, v. 8, n. 1, p. 1-37, 2021.

E-PRINTS IN LIBRARY AND INFORMATION SCIENCE (E-LIS). **Quem somos?** Nápoles, 2022. Disponível em: <http://eprints.rclis.org/privacy.html>. Acesso em: 18 jun. 2022.

FELICETTI, Vera Lúcia *et al.* Perspectivas emergentes en investigación educativa. **Revista Tempos e Espaços em Educação**, v. 12, n. 30, p. 155-170, 2019. Disponível em: <https://seer.ufs.br/index.php/revtee/article/view/10681/pdf>.

FERNANDES, Florestan. **A universidade brasileira: reforma ou revolução?**. São Paulo: Ed. Alfa - Ômega, 1975. p. 272

FERNANDES, Henrique Denes Hilgenberg; VILAN FILHO, Jayme Leiro. Fluxo da informação científica: uma revisão dos modelos propostos na literatura em Ciência da Informação. **Em Questão**, Porto Alegre, v. 27, n. 2, p. 138-154, abr./jun. 2021. Disponível em: <https://www.redalyc.org/journal/4656/465666428007/465666428007.pdf>. Acesso em: 20 abr. 2022.

FERREIRA, João Batista; SILVA, Luciana de Araújo Mendes. O uso da bibliometria e sociometria como diferencial em pesquisas de revisão. **Revista Brasileira de Biblioteconomia e Documentação**, São Paulo, v. 15, n. 2, p. 448-464, maio/ago. 2019. Disponível em: <https://11nq.com/XXguE>. Acesso em: 29 jul. 2022.

FONSECA, Edson Nery da (Org.). **Bibliometria: teoria e prática**. São Paulo: Editorada USP, 1986.

FROHMANN, Bernd. The power of images: a discourse analysis of the cognitive viewpoint. **Journal of Documentation**, [S. l.], v. 48, n. 4, p. 365-386, 1992.

FÜHR, Fabiane; ALVAREZ, Edgar Bisset. Produção científica sobre humanidades digitais e ciência aberta na Web da Ciência (WOS) e bancos de dados Scopus - 2014-2020. **Anales de Investigación**, Madri, v. 17, n. 3, p. 1-24, set./dez. 2021. Disponível em: <https://docs.google.com/viewerng/viewer?url=http://revistas.bnjm.cu/index.php/BAI/article/viewFile/441/412>. Acesso em: 26 abr. 2022.

FURCHE, Tim *et al.* Data Wrangling for Big Data: Challenges and Opportunities. In: **EDBT**. 2016. p. 473-478.

FUNDAÇÃO CASA DE RUI BARBOSA (FCRB). Programa de Pós-Graduação em Memória e Acervos. **Apresentação**. Rio de Janeiro, 2022. Disponível em: <http://mestrado.casaruibarbosa.gov.br/>. Acesso em: 27 maio 2022.

FURTADO, Caio Alves. **Análise espaço-temporal da trajetória de carreira de pesquisadores brasileiros**. 2015. 120f. Dissertação (Mestrado em Ciência da Computação) – Instituto de Ciências Exatas, Universidade Federal de Minas Gerais, Belo Horizonte, 2015. Disponível em: <https://l1nq.com/1UPpz>. Acesso em: 20 abr. 2022.

FRAZÃO, Ana.; OLIVA, Milena Donato; TEPEDINO, Gustavo. **Lei geral de proteção de dados pessoais e suas repercussões no direito brasileiro**. São Paulo: Thomson Reuters Brasil, 2019.

FREITAS, Marília Augusta de; LEITE, Fernando César Lima. Atores do sistema de comunicação científica: apontamentos para discussão de suas funções. **Informação & Informação**, Londrina, v. 24, n. 1, p. 273-299, jan./abr. 2019. Disponível em: <https://repositorio.unb.br/handle/10482/34121>. Acesso em: 19 abr. 2022.

GABRIEL JUNIOR, Rene Faustino; BOCHI, Fernanda; MOURA, Ana Maria Mielniczuk de. Aproximações da produção científica em ciências da saúde na ciência da informação no Brasil. **Revista Eletrônica de Comunicação, Informação & Inovação em Saúde**, Rio de Janeiro, v. 15, n. 4, p. 824-839, out./dez. 2021. Disponível em: <https://encr.pw/blxt1>. Acesso em: 30 jul. 2022.

GALDINO, Rosangela; GARCIA, Leonardo Guimarães; AMARAL, Roniberto Morato do. Contribuições da bibliometria ao tratamento de dados institucionais não sistematizados de produção científica: o caso do Instituto Federal de Educação, Ciência e Tecnologia de São Paulo (IFSP). **Brazilian Journal of Information Science: Research trends**, Marília, v. 16, n. 2, p. 1-21, jul./dez. 2022. Disponível em: <https://l1nq.com/lby3y>. Acesso em: 30 jul. 2022.

GALVÃO, Kécia da Silveira; FERREIRA, Gustavo Henrique de Aragão; ANDRADE, Cacilda Soares de. Tutoria e educação a distância: uma investigação bibliométrica a partir da mineração de texto. In: CONGRESSO INTERNACIONAL DE EDUCAÇÃO E TECNOLOGIAS, 1., 2020, São Carlos. **Anais [...]**. São Carlos: UFSCar, 2020. p. 1-19. Disponível em: <https://encr.pw/45KZ9>. Acesso em: 30 jul. 2022.

GAMA, Jean Carlos Freitas; FERREIRA NETO, Amarílio; SANTOS, Wagner dos. Formação para atuação com o esporte: características bibliométricas e redes de colaboração. **Movimento**, Porto Alegre, v. 27, n. 1, p. 1-19, jan./jun. 2021. Disponível em: <https://l1nq.com/6crjF>. Acesso em: 29 jul. 2022.

GARVEY, William. **Communication: the essence of Science**. Oxford: Pergamon, 1979.

GEOCAPES. **Dados estatísticos**. [S. l.], 2022. Disponível em: <https://geocapes.capes.gov.br/geocapes/>. Acesso em: 20 abr. 2022.

GIL, Antônio Carlos. **Como elaborar projetos de pesquisa**. 4. ed. São Paulo: Atlas, 2002.

GOMES, Henriette Ferreira. Interdisciplinaridade e Ciência da Informação: de característica a critério delineador de seu núcleo principal. **DataGramZero: Revista de ciência da informação**, v. 2, n. 4, p. 1-8, 2001. Acesso em: 20 abr. 2022.

GONZÁLEZ-ALCAIDE, Gregorio. Bibliometric studies outside the information science and library science field: uncontainable or uncontrollable?. **Scientometrics**, v. 126, n. 8, p. 6837-6870, 2021.

GOOGLE. Google Maps. Disponível em: <https://www.google.com.br/maps>. Acesso em: 03 de março de 2025.

GONÇALVES, André Vinícius; KARPINSKI, Cezar; ARAÚJO, Gustavo Medeiros de. Perfil acadêmico interdisciplinar dos pesquisadores da Ciência da Informação do Brasil: indicadores da Plataforma Lattes. **Brazilian Journal of Information Studies: Research trends**, Marília, v. 15, n. 1, p. 1-20, jan. 2021. Disponível em: <https://encr.pw/AL5nr>. Acesso em: 20 abr. 2022.

GONTIJO, Marília Catarina Andrade; ARAÚJO, Ronaldo Ferreira de. Impacto acadêmico e atenção on-line de pesquisas sobre inteligência artificial na área da saúde: análise de dados bibliométricos e altmétricos. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, Florianópolis, v. 26, n. 1, p. 1-21, jan. 2021. Disponível em: <https://encr.pw/5ldM2>. Acesso em: 26 abr. 2022.

GRÁCIO, Maria Cláudia Cabrini. **Análises relacionais de citação para a identificação de domínios científicos**: uma aplicação no campo dos Estudos Métricos da Informação no Brasil. São Paulo: Editora UNESP, 2020.

GRANDVALET, Yves; BENGIO, Yoshua. Semi-supervised learning by entropy minimization. **Advances in neural information processing systems**, v. 17, 2004.

GRAVES, Alvaro; HENDLER, James. A study on the use of visualizations for open government data. **Information Polity**, v. 19, n. 1-2, p. 73-91, 2014.

GROGAN, Denis Joseph. **A prática do serviço de referência**. Tradução de Antônio Agenor Briquet Lemos. Brasília-DF: Briquet Lemos, 2001.

GUEDES, Vânia; BORSCHIVER, Suzana. Bibliometria: uma ferramenta estatística para a gestão da informação e do conhecimento em sistemas de informação, de comunicação e de avaliação científica e tecnológica. In: ENCONTRO NACIONAL DE PESQUISA E PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO, 6., 2005, Salvador. **Anais [...]**. Salvador: UFBA, 2005. p. 1-18. Disponível em: <https://encr.pw/Qjecx>. Acesso em: 25 abr. 2022.

GUIMARÃES, Tomás Aquino *et al.* A ANPAD e o processo de institucionalização da comunidade científica brasileira de Administração. **Cadernos EBAPE.BR**, Rio de Janeiro, v. 16, n. esp., p. 523-537, out. 2018. Disponível em: <https://encr.pw/JkNPO>. Acesso em: 20 abr. 2022.

GUIMARÃES, André José Ribeiro; BEZERRA, Cícero Aparecido. Gestão de dados: uma análise bibliométrica. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 24, n. 4, p. 171-186, out./dez. 2019. Disponível em: <https://11nq.com/fWIL4>. Acesso em: 25 abr. 2022.

GREGORI, Maria Stella. Os impactos da Lei Geral de Proteção de Dados Pessoais na saúde suplementar. **Revista de Direito do Consumidor**, p. 171-196, 2020. Acesso em: 05 nov. 2022.

GUNAWAN, Teddy Surya *et al.* Development of video-based emotion recognition using deep learning with Google Colab. **Telecommunication Computing Electronics and Control**, [S. l.], v. 18, n. 5, p. 2463-2471, maio 2020. Disponível em: <https://encr.pw/OA8HP>. Acesso em: 22 out. 2022.

HAYASHI, Maria Cristina Piumbato Innocentini; RIGOLIN, Camila Carneiro Dias; HAYASHI, Carlos Roberto Massao. Transgênicos e sociedade na agenda das pesquisas acadêmicas brasileiras. **Política & Sociedade**, v. 11, n. 20, p. 99-134, 2012.

HAM, Kelli. OpenRefine (version 2.5): free, open-source tool for cleaning and transforming data. **Journal of the Medical Library Association**, [S. l.], v. 101, n. 3, p. 233-234, mar. 2013. Disponível em: <https://11nq.com/XEDB2>. Acesso em: 03 nov. 2022.

HEER, Jeffrey; SHNEIDERMAN, Ben. Interactive dynamics for visual analysis: A taxonomy of tools that support the fluent and flexible use of visualizations. **Queue**, v. 10, n. 2, p. 30-55, 2012.

HJØRLAND, Birger. Epistemology and the socio-cognitive perspective in Information Science. **Journal of the American Society for Information Science and Technology**, [S. l.], v. 53, n. 4, p. 257-270, 2002.

HJØRLAND, Birger. Information science and its core concepts: Levels of disagreement. In: **Theories of information, communication and knowledge: A multidisciplinary approach**. Dordrecht: Springer Netherlands, 2013. p. 205-235.

HOSTINS, Regina Célia Linhares. Os Planos Nacionais de Pós-graduação (PNPG) e suas repercussões na Pós-graduação brasileira. **Perspectiva**, v. 24, n. 1, p. 133-160, 2006.

HU, Haotian; WANG, Dongbo; DENG, Sanhong. Global Collaboration in Artificial Intelligence: Bibliometrics and Network Analysis from 1985 to 2019. **Journal of Data and Information Science**, [S. l.], v. 5, n. 4, p. 86-115, abr. 2020. Disponível em: <https://sciendo.com/es/article/10.2478/jdis-2020-0027>. Acesso em: 26 abr. 2022.

HUANG, Chih-Wei *et al.* A richly interactive exploratory data analysis and visualization tool using electronic medical records. **BMC medical informatics and decision making**, v. 15, p. 1-14, 2015.

INFORMATION SCIENCE & TECHNOLOGY ABSTRACTS (ISTA). **Sobre a Ebsco**. [S. l.], 2022. Disponível em: <https://encr.pw/xPz0j>. Acesso em: 18 jun. 2022.

INSTITUTO DE PESQUISA ECONÔMICA APLICADA (Ipea). **Nordeste poderia crescer mais que o Brasil até 2030**. Publicado em 14 set. 2021. Disponível em: <https://encurtador.com.br/1A8gV>. Acesso em: 30 mar. 2025

IVO, Naiara Vieira Silva *et al.* Análise da produção científica sobre política de informação no Brasil na Plataforma Lattes. **Anales de Investigación**, Madri, v. 17, n. 2, p. 82-101, maio/ago. 2021. Disponível em: <https://11nq.com/AZtw4>. Acesso em: 25 abr. 2022.

ISLAM, Mohaiminul; JIN, Shangzhu. An overview of data visualization. In: 2019 INTERNATIONAL CONFERENCE OF INFORMATION AND COMMUNICATIONS

TECHNOLOGIES (ICISCT). Tashkent, Uzbekistan, 2019, p. 1-7. DOI 10.1109/ICISCT47635.2019.9012031.

JOLY, Maria Cristina Rodrigues Azevedo *et al.* Análise de teses e dissertações em avaliação psicológica disponíveis na BVS-PSI Brasil. **Psicologia: ciência e profissão**, v. 30, p. 174-187, 2010.

JÚNIOR, Jaime Magalhães Sepulcro; GALVÃO, Ana Maria de Oliveira. protestantismo, história e cultura escrita: um balanço da produção discente na pós-graduação brasileira. **Comunicações**. v. 25, n. 3, p. 223-245. 2018. DOI: <https://11nq.com/Sp76a>

JUSTINO, Thais da Silva. **Análise da Colaboração Científica dos Programas de Pós-Graduação em Ciência da Informação Brasileiros**. 2019. 103f. Dissertação (Mestrado em Ciência da Informação) – Centro de Informação e Comunicação, Universidade Federal de São Carlos, São Carlos, 2019. Disponível em: <https://encurtador.com.br/JmZnf>. Acesso em: 20 abr. 2022.

KAWALEC, Anna. Research trends in library and information science based on Spanish scientific publication 2000 to 2010. **Malaysian journal of library and information science**, v. 18, n. 2, p. 1-13, 2013.

KITCHIN, Rob. **The data revolution: Big data, open data, data infrastructures and their consequences**. Califórnia: Sage, 2014.

KURTULUŞ, Muhammed Akif; TATAR, Nilgün. An Analysis of Scientific Articles on Science Misconceptions: a Bibliometric Research. **Ilkogretim Online: Elementary Education Online**, [S. l.], v. 20, n. 1, p. 192-207, jan./jun. 2021. Disponível em: <https://encurtador.com.br/nblzP>. Acesso em: 25 abr. 2022.

KUSUMASARI, Tien Fabrianti *et al.* Data profiling for data quality improvement with OpenRefine. In: 2016 INTERNATIONAL CONFERENCE ON INFORMATION TECHNOLOGY SYSTEMS AND INNOVATION (ICITSI). Bandung, Indonésia, 2016, p. 1-6. DOI: 10.1109/ICITSI.2016.7858197..

LANÇA, Tamie Aline; AMARAL, Roniberto Morato; GRACIOSO, Luciana Souza. Multi e interdisciplinaridade nos programas de pós-graduação em Ciência da Informação brasileiros. **Perspectivas em Ciência da Informação**, v. 23, p. 150-183, 2018.

LE COADIC, Yves-François. **A Ciência da Informação**. 2. ed. Brasília: Briquet de Lemos, 2004.

LE COADIC, Yves-François. **A Ciência da Informação**. Tradução de Maria Yêda F. S. de Filgueiras Gomes. Brasília: Briquet de Lemos, 1996.

LENZERINI, Maurizio. Data integration: A theoretical perspective. In: ACM SIGMOD-SIGACT-SIGART SYMPOSIUM OF PRINCIPLES OF DATABASE SYSTEMS. **Proceedings** [...]. Madison, Wisconsin, USA, 2002. p. 233-246. DOI:10.1145/543613.543644

LI, Michael Lingzhi. **Algorithms for Large-scale Data Analytics and Applications to the COVID-19 Pandemic**. 2022. Tese (Doutorado) - Massachusetts Institute of Technology, Cambridge/USA, 2022.

LIBRARY AND INFORMATION SCIENCE ABSTRACTS (LISA). **Sobre a ProQuest**. [S. l.], 2022. Disponível em: <https://www.proquest.com/>. Acesso em: 18 jun. 2022.

LIBRARY, INFORMATION SCIENCE & TECHNOLOGY ABSTRACTS (LISTA). **Uma base de dados gratuita de pesquisa para estudos em Biblioteconomia e Ciência da Informação**. EBISCO, [S. l.], 2022. Disponível em: <https://l1nq.com/uc3W7>. Acesso em: 18 jun. 2022.

LIMA, Leonia Maria de *et al.* Palliative care in primary health care: scientific production of theses and dissertations in the brazilian framework. **Revista de Pesquisa, Cuidado é Fundamental Online**, v. 13, p. 1362-1367, 2021.

LIMA, P. R. S.; FERREIRA, J. R. S.; PRADO, M. A. R. Produções científicas nos subcampos dos estudos métricos da informação. In: ENCONTRO BRASILEIRO DE BIBLIOMETRIA E CIENTOMETRIA, 08., 2022. **Anais [...]** VIII Encontro Brasileiro de Bibliometria e Cientometria, 2022. DOI: <https://doi.org/10.22477.52>

LIN, Jimmy *et al.* Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In: 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2021. **Proceedings [...]** p. 2356-2362. DOI: <https://doi.org/10.1145/3404835.3463238>

LITTLE, Roderick J. *et al.* The prevention and treatment of missing data in clinical trials. **New England Journal of Medicine**, v. 367, n. 14, p. 1355-1360, 2012.

LOPES, Gesiel Rios *et al.* Introdução à análise exploratória de dados com python. **Minicursos ERCAS ENUCMPI**, v. 2019, p. 160-176, 2019.

LOPES, Pâmela Tamires Dias; AGANETTE, Elisângela Cristina; MACULAN, Benildes Coura. Taxonomia corporativa e taxonomia facetada: usos e aplicações na ciência da informação no Brasil. **Investigación bibliotecológica**, v. 34, n. 82, p. 159-173, 2020.

LUNA, Débora de Oliveira Lopes do Rego; LUNA, Amanda Souza Xavier de. Plano Nacional de Pós-Graduação 2011-2020: metas e desafios para o desenvolvimento da Pós-Graduação no Brasil. In: COLÓQUIO INTERNACIONAL DE PESQUISAS EM EDUCAÇÃO SUPERIOR, 4., 2015, João Pessoa. **Anais [...]**. João Pessoa: UFPB, 2017. p. 1-16. Disponível em: <https://l1nq.com/cdl5P>. Acesso em: 20 abr. 2022.

LUNARDELLI, Rosane Suely Alvares; PAIVA, Andréia Del Conte de; KAWAKAMI, Tatiana Tissa. Os resumos e suas especificidades em periódicos brasileiros da ciência da informação. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, Florianópolis, v. 25, n. 2, p. 1-17, jul./dez. 2020. Disponível em: <https://l1nq.com/Tigky>. Acesso em: 24 abr. 2022.

MACHADO, Dayan Eduardo Aguilera. Compendio de hojas de cálculo con LibreOffice Calc para el procesamiento de información estadística. **Revista Cubana de Tecnología de la Salud [Internet]**, v. 15, n. 2, p. 4298, 2024.

MACHADO JÚNIOR, Celso *et al.* As leis da bibliometria em diferentes bases de dados. **Revista de Ciências da Administração**, [S. l.], v. 18, n. 44, p. 111-123, abr. 2016. Disponível em: <https://11nq.com/bmqxj>. Acesso em: 25 abr. 2022.

MAIA, Manuela Eugênio *et al.* Ciência da Informação em foco: análise bibliométrica da produção científica da Revista Analisando em Ciência da Informação. **RACIn**, João Pessoa, v. 8, n. 1, p. 45-60, jan./jun. 2020. Disponível em: <https://encr.pw/ROT8Z>. Acesso em: 25 abr. 2022.

MARCONI, Marina Andrade; LAKATOS, Eva Maria. **Técnicas de pesquisa**. 9. ed. São Paulo: Atlas, 2021.

MAROLDI, Alexandre Masson *et al.* Panorama bibliométrico das teses e dissertações sobre educação indígena. **Revista Brasileira de Educação do Campo**, v. 2, n. 2, p. 677-707, 2017.

MARQUES, Francis Bento; MACULAN, Benildes Coura Moreira dos Santos; SOUZA, Renato Rocha. Evaluating FAIR principles compliance in the Capes Open Data Portal: a case study of Open Government Data. **Brazilian Journal of Information Science**, n. 18, p. 35, 2024.

MATA, Marta Leandro da. Contribuições dos estudos acerca da competência em informação para a ciência da informação: uma análise a partir da produção científica do Enancib entre 2015 a 2019. **Informação & Informação**, Londrina, v. 26, n. 1, p. 232-263, jan./jun. 2021. Disponível em: <https://11nq.com/Ij1K8>. Acesso em: 25 abr. 2022.

MARTELETO, Regina Maria. A pesquisa em Ciência da Informação no Brasil: marcos institucionais, cenários e perspectivas. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 14, n. esp., p. 19-40, jan. 2009. Disponível em: <https://11nq.com/d9GaX>. Acesso em: 24 abr. 2022.

MARTINS, Ana Carolina de Melo *et al.* Mestrado Profissional na área de Ciência da Informação no Brasil: o caso de Santa Catarina. **Revista ACB**, Florianópolis, v. 20, n. 3, p. 411-422, set./dez. 2015. Disponível em: <https://encr.pw/LgM2h>. Acesso em: 27 maio 2022.

MATTEDI, Marcos Antônio; SPIESS, Maiko Rafael. Modalidades de regulação da atividade científica: uma comparação entre as interpretações normativa, cognitiva e transacional dos processos de integração social da comunidade científica. **Educação Social**, Campinas, v. 31, n. 110, p. 73-92, jan./mar. 2010. Disponível em: <https://encr.pw/59t8n>. Acesso em: 20 abr. 2022.

MCKINNEY, Wes. **Python for data analysis: Data Wrangling With Pandas, NumPy and IPython**. [S. l.]: O'Reilly Media, 2012.

MEADOWS, Arthur Joseph. **A comunicação científica**. Brasília: Briquet de Lemos, 1999.

MEDEIROS, Wagner Oliveira de; PINHO, Fabio Assis; CORREA, Renato Fernandes. Aplicação de *software* de mineração de texto na representação da informação de obras artístico-pictóricas. **Logeion: Filosofia da informação**, Rio de Janeiro, v. 6, n. 1, p.146-170, fev. 2020. Disponível em: <https://11nq.com/Txblb>. Acesso em: 26 abr. 2022.

MEDEIROS, José Mauro Gouveia de; VITORIANO, Maria Albeti Vieira. A evolução da bibliometria e sua interdisciplinaridade na produção científica brasileira. **Revista Digital de Biblioteconomia e Ciência da Informação**, Campinas, v. 13, n. 3, p. 491-503, set./dez. 2015. Disponível em: <https://11nq.com/Pe3nc>. Acesso em: 30 jul. 2022.

MELO, João Henrick Neri de; TRINCA, Tatiane Pacanaro; MARICATO, João de Melo. Limites dos indicadores bibliométricos de bases de dados internacionais para avaliação da Pós-Graduação brasileira: a cobertura da Web of Science nas diferentes áreas do conhecimento. **TransInformação**, Campinas, v. 33, n. 1, p. 1-16, jan./jun. 2021. Disponível em: <https://11nq.com/Wet0g>. Acesso em: 20 abr. 2022.

MERTON, Robert K. The Matthew effect in science: the reward and communication systems of science are considered. **Science**, [S. l.], v. 159, n. 3810, p. 56–63, 1968. Disponível em: <https://11nq.com/EMY3u>. Acesso em: 11 abr. 2025.

MONTGOMERY, Scott L. **Does science need a global language?:** English and the future of research. Chicago: University of Chicago Press, 2013.

MORAES, Ivan Furegato; AMARAL, Cacilda Mendes dos Santos; BASTOS, Flávia da Cunha. Teses de doutorado em gestão do esporte no Brasil: uma revisão integrativa metodológica. **Movimento**, v. 27, p. e27012, 2021.

MORAES, Lena Lúcia de; KAFURE, Ivete. Bibliometria e ciência de dados: um exemplo de busca e análise de dados da Web of Science (WoS). **Revista Digital de Biblioteconomia e Ciência da Informação**, Campinas, v. 18, n. 1, p. 1-20, jan. 2020. Disponível em: <https://11nq.com/MF3Pg>. Acesso em: 26 abr. 2022.

MOREIRA, Paulo Sérgio da Conceição; GUIMARÃES, André José Ribeiro; TSUNODA, Denise. Qual ferramenta bibliométrica escolher? Um estudo comparativo entre *softwares*. **Revista P2P e INOVAÇÃO**, [S. l.], v. 6, n. 1, p. 140-158, jan. 2020. Disponível em: <https://brapci.inf.br/index.php/res/v/137643>. Acesso em: 26 abr. 2022.

MOREIRA, Jonathan Rosa; FERNEDA, Edberto. Produção científica nos programas de pós-graduação nas áreas de informação no Brasil. **Informação & Informação**, v. 25, n. 4, p. 142-168, 2020.

MOROSINI, Marília Costa. A Pós-Graduação no Brasil. **Revista Argentina de Educação Superior**, Buenos Aires, v. 18, n. 1, p. 125-152, jan./jun. 2009. Disponível em: <https://encr.pw/COgFz>. Acesso em: 20 abr. 2022.

MOTIN, Sirlene Donaiski *et al.* Educação ambiental na formação inicial docente: um mapeamento das pesquisas brasileiras em teses e dissertações. **Investigações em Ensino de Ciências**, v. 24, n. 1, 2019.

MUELLER, Suzana P. M. A Ciência, o sistema de comunicação científica e a literatura científica. In: CAMPELLO, Bernadete Santos *et al.* (org). **Fontes de informação para pesquisadores e profissionais**. Belo Horizonte: UFMG, 2000. p. 21-34.

NASCIMENTO, Mateus Rebouças; PINTO, Adilson Luiz; DIAS, Thiago Magela Rodrigues. Análise da produção intelectual na Pós-Graduação em Ciência da Informação: um estudo bibliométrico baseado em dados da Plataforma Lattes. **Biblion Analis Investigacion**, [S. l.], v. 16, n. 3, p. 207-220, mar. 2020. Disponível em: <https://encr.pw/1ix0q>. Acesso em: 20 abr. 2022.

NOBRE, Lorena Neves; FREITAS, Rodrigo. A evolução da Pós-Graduação no Brasil: histórico, políticas e avaliação. **Brazilian Journal of Production Engineering**, São Mateus, v. 3, n. 2, p. 18-30, jul./dez. 2017. Disponível em: <https://encr.pw/m7V6i>. Acesso em: 20 abr. 2022.

NONGTHOMBAM, Kiranbala; SHARMA, Deepika. Data Analysis using Python. **International Journal of Engineering Research & Technology (IJERT)**, v. 10, n. 7, 2021.

NANDA, Santosh *et al.* Emphasis on validation in research: A meta-analysis. **Scientometrics**, v. 48, n. 1, p. 45-64, 2000.

OLIVEIRA, Carlos Alexandre de. **A genealogia acadêmica da ciência da informação brasileira: análise dos currículos dos pesquisadores/docentes**. 2021. 171f. Tese (Doutorado em Gestão e Organização do Conhecimento) – Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2021. Disponível em: <https://encr.pw/42uJQ>. Acesso em: 20 abr. 2022.

OLIVEIRA, Edivaldo Herculano Corrêa de. Redes de colaboração em pesquisa e intercâmbio de conhecimento científico. **Revista Pan-Amazônica em Saúde**, [S. l.], v. 9, n. 4, p. 9-11, abr. 2018. Disponível em: <https://11nk.dev/LGn5A>. Acesso em: 26 abr. 2022.

OLIVEIRA, Kalyane Kelly Duarte de; MAIA, Clécio André Alves da Silva; MIRANDA, Francisco Arnaldo Nunes de. Teses e dissertações em enfermagem baseadas na teoria das representações sociais: análise bibliométrica. **Revista de Pesquisa, Cuidado é Fundamental Online**, v. 8, n. 1, p. 4079, 2016.

ORTEGA, Fábio Silva; BRANDÃO, Carlos da Fonseca. A história da Pós-Graduação no Brasil e a construção do espaço acadêmico científico da Educação. **Educação em Foco**, [S. l.], v. 23, n. 39, p. 249-269, jan./abr. 2020. Disponível em: <https://encr.pw/oHQXO>. Acesso em: 20 abr. 2022

PAIM, Isis *et al.* Interdisciplinaridade na ciência da informação: início de um diálogo. **Perspectivas em ciência da informação**, v. 6, n. 1, 2001

PAIVA, Thalís Regina Silva *et al.* Scientific production of dissertations and theses on palliative care and chronic diseases: Bibliometric study. **Revista De Pesquisa, Cuidado é Fundamental Online**, v. 12, p. 723-729, 2020.

PAZOS, Diego Marques; SIRQUEIRA, Tássio Ferenzini Martins. Análise de Dados em Gráficos com base na API do Twitter. **Caderno de Estudos em Engenharia de Software**, v. 3, n. 2, 2022.

PEREIRA, Raquel *et al.* Metanálise como instrumento de pesquisa: uma revisão sistemática dos estudos bibliométricos em administração. **RAM**, São Paulo, v. 20, n. 5, p. 1-33, maio 2019. Disponível em: <https://encr.pw/wiXvF>. Acesso em: 30 jul. 2022.

PENG, Chao-Ying Joanne; LEE, Kuk Lida; INGERSOLL, Gary M. An introduction to logistic regression analysis and reporting. **The journal of educational research**, v. 96, n. 1, p. 3-14, 2002.

PIERSON, Lillian. **Data science for dummies**. Nova Jersey: John Wiley & Sons, 2021.

PILATI, Orlando. Especialização: falácia ou conhecimento aprofundado? **Revista Brasileira de Pós-Graduação**, [S. l.], v. 3, n. 5, p. 7-26, jun. 2006.

PINHEIRO, Lena Vania Ribeiro. Infra-estrutura da pesquisa em Ciência da Informação no Brasil. **DataGramZero**. Rio de Janeiro: IASI, v.1, n.6, dez. 2000.

PINHEIRO, Lena Vânia Ribeiro. Confluências interdisciplinares entre Ciência da Informação e Museologia. **Revista Museologia & Interdisciplinaridade**, v. 1, n. 1, p. 07-31, 2012.

PINTO, Lourival. Significados e contextos: leitura e interpretação na ciência da informação. **Revista Digital Biblioteconomia e Ciência da Informação**, Campinas, v. 16, n. 3, p. 338-354, set./dez. 2018. Disponível em: <https://11nq.com/emE1u>. Acesso em: 24 abr. 2022.

PINTO, Maria *et al.* Scientific production on mobile information literacy in higher education: a bibliometric analysis (2006–2017). **Scientometrics**, [S. l.], v. 120, n. 1, p. 57-85, jan./jun. 2019. Disponível em: <https://11nq.com/INaIM>. Acesso em: 25 abr. 2022.

PIURCOSKY, Fabrício Peloso *et al.* A lei geral de proteção de dados pessoais em empresas brasileiras: uma análise de múltiplos casos. **Suma de negócios**, v. 10, n. 23, p. 89-99, 2019.

PRADHAN, Pallab. Scientific mapping and visualization tools used in bibliometric and scientometric studies: an overview. **Inflibnet**, [S. l.], v. 23, n. 4, p. 19-33, out./dez. 2017. Disponível em: <https://ir.inflibnet.ac.in:8443/ir/handle/1944/2132>. Acesso em: 26 abr. 2022.

PRIYA, Piyush *et al.* Stress combinations and their interactions in plants database: a one-stop resource on combined stress responses in plants. **The Plant Journal**, v. 116, n. 4, p. 1097-1117, 2023.

PRICE, Derek. Networks of scientific papers. **Science**, [S. l.], v. 149, n. 7, p. 510-515, jul. 1965.

PRIETO-GUTIÉRREZ, Juan José; SEGADO-BOJ, Francisco. Annals of library and information studies: a bibliometric analysis of the journal and a comparison with the top library and information studies journals in Asia and worldwide (2011–2017). **The Serials Librarian**, v. 77, n. 1-2, p. 38-48, 2019.

QUEIROZ, Fernanda Mendes; NORONHA, Daisy Pires. Temática das dissertações e teses em ciência da informação no Programa de Pós-Graduação em Ciências da Comunicação da USP. **Ciência da Informação**, v. 33, p. 132-142, 2004.

QUEIROZ, Daniela Gralha de Caneda; VILAN FILHO, Jayme Leiro; MOURA, Ana Maria Mielniczuk de. Visibilidade de artigos de periódicos científicos brasileiros de matemática nas bases de dados internacionais. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 16., 2015, João Pessoa. **Anais...** João Pessoa: ANCIB, 2015. Disponível em <http://www.ufpb.br/evento/lti/ocs/index.php/enancib2015/enancib2015/paper/viewFile/2966/1142>. Acesso em: 26 maio 2022.

RAHM, Erhard *et al.* Data cleaning: Problems and current approaches. **IEEE Data Eng. Bull.**, v. 23, n. 4, p. 3-13, 2000.

RAPÔSO, Cláudio Filipe Lima *et al.* Lgpd-lei geral de proteção de dados pessoais em tecnologia da informação: Revisão sistemática. **RACE-Revista de Administração do Cesmac**, v. 4, p. 58-67, 2019.

RIBEIRO, Kíssila da Conceição *et al.* Indicadores bibliométricos na base Scopus sobre o tema “avaliação da educação” no Brasil e no mundo. **Revista Tecnologia e Sociedade**, [S. l.], v. 15, n. 37, p. 311-331, dez. 2019. Disponível em: <https://revistas.utfpr.edu.br/rts/article/view/8436/6310>. Acesso em: 30 jul. 2022.

RIOS, Francisco Welton Silva; SOUZA, Maria Naires Alves de. Gênero e sexualidade como temas de teses e dissertações: levantamento quantitativo nos repositórios do IBICT e da CAPES. **Revista Brasileira de Biblioteconomia e Documentação**, v. 13, n. esp. CBBD, p. 1841-1856, 2017.

RODRIGUES, Carlos Manoel Lopes; FAIAD, Cristiane. Pesquisa sobre riscos psicossociais no trabalho: estudo bibliométrico da produção nacional de 2008 a 2017. **Revista Psicologia: Organizações e Trabalho**, [S. l.], v. 19, n. 1, p. 571-579, jan./mar. 2019. Disponível em: <https://encr.pw/r7jxx>. Acesso em: 29 jul. 2022.

RODRIGUES, Maria Rosemary; SERVANTES, Brígida Maria Nogueira. Os mapas conceituais para a visualização de conceitos de áreas do conhecimento em unidades de informação. **Revista ACB: Biblioteconomia em Santa Catarina, Florianópolis**, v. 18, n. 1, p. 752-776, jan./jun. 2013. Disponível em: <https://encr.pw/acaet>. Acesso em: 29 jul. 2022.

ROZA, Rodrigo Hipólito. Ciência da Informação, tecnologia e sociedade. **Biblos**, Rio Grande, v. 32, n. 2, p. 177-190, jul./dez. 2018. Disponível em: <https://seer.furg.br/biblos/article/view/7546/5861>. Acesso em: 27 maio 2022.

SÁ-SILVA, Jackson Ronie *et al.* Pesquisa documental: pistas teóricas e metodológicas. **Revista brasileira de história & ciências sociais**, v. 1, n. 1, p. 1-15, 2009. Disponível em: <https://periodicos.furg.br/rbhcs/article/view/10351/pdf>. Acesso em: 14 mar. 2025.

SANTANA, Guilherme Alves de. **A produção colaborativa de conhecimento dos grupos de pesquisa brasileiros e os desdobramentos das relações entre seus pesquisadores**. 2015. 165f. Dissertação (Mestrado em Ciência da Informação) – Centros de Artes e Comunicação,

Universidade Federal de Pernambuco, Recife, 2015. Disponível em: <https://encr.pw/d32GO>. Acesso em: 20 abr. 2022.

SANTOS, Marcio Antonio Raiol; PEREIRA, Humberto. Análise da produção acadêmica sobre a ludicidade na prática docente: o ensino fundamental como etapa emergente. **Dialogia**, São Paulo, v. 31, n. 1, p. 45-56, jan./abr. 2019. Disponível em: <https://periodicos.uninove.br/dialogia/article/view/11454>. Acesso em: 20 jun. 2022.

SARACEVIC, Tefko. The interdisciplinary nature of information science. **Ciência da Informação**, Brasília, v. 24, n. 1, p. 36-41, jan./abr. 1995.

SARACEVIC, Tefko. Ciência da informação: origem, evolução e relações. **Perspectivas em ciência da informação**, v. 1, n. 1, 1996.

SARACEVIC, Tefko. Digital library evaluation: Toward an evolution of concepts. **Library Trends**, v. 49, n. 2, p. 350-69, 2000.

SARACEVIC, Tefko. Information Science. In: BATES, Marcia; MAACK, Mary Niles (eds.). **Encyclopedia of Library and Information Science**. New York: Taylor & Francis, 2009. p. 2570-2586.

SARTI, Thiago Dias. O desafio da qualidade e integridade das publicações científicas. **Revista Brasileira de Medicina de Família e Comunidade**, v. 12, n. 39, p. 1-4, 2017.

SCHWEITZER, Fernanda; RODRIGUES, Rosângela Schwarz. Produção científica em áreas multidisciplinares: educação a distância no Brasil. **Em Questão**, Porto Alegre, v. 19, n. 1, p. 156-172, jan./jun. 2013. Disponível em: <http://seer.ufrgs.br/index.php/EmQuestao/article/view/30924>. Acesso em: 25 maio 2023.

SCOPUS. **O que é Scopus?** [S. l.], 2022. Disponível em: https://www-elsevier-com.ez27.periodicos.capes.gov.br/solutions/scopus?dgcid=RN_AGCM_Sourced_300005030. Acesso em: 18 jun. 2022.

SHARMA, Jyoti; CHAKRAVARTY, Rupak. Application of Lotka's Law in Library Science Literature of Select Central Universities in North India. **Journal of Advances in Library and Information Science**, [S. l.], v. 7, n. 1, p. 36-39, jan./jun. 2018. Disponível em: <http://jalis.in/pdf/7-1/Jyoti.pdf>. Acesso em: 25 abr. 2022.

SHERA, Jesse H. Epistemologia social, semântica e biblioteconomia. **Ciência da Informação**, Brasília, v. 1, p. 7-16, 1977. Disponível em: <https://revista.ibict.br/ciinf/article/view/92>. Acesso em: 12 abr. 2025.

SEMBAY, Márcio José *et al.* Aplicação da Lei de Bradford a pesquisas relacionadas a open government. **Anales de Documentación**, Madri, v. 23, n. 1, p. 1-10, jan./jun. 2020. Disponível em: <https://revistas.um.es/analesdoc/article/view/326771/274341>. Acesso em: 25 abr. 2022.

SIDDIQUE, Nadeem; REHMAN, Shafiq; KHAN, Muhammad. Library research and information science in Pakistan: a bibliometric analysis, 1957-2018. **Journal of Library and**

Information Science, [S. l.], v. 53, n. 1, p. 89-102, jan./jun. 2021. Disponível em: <https://journals.sagepub.com/doi/full/10.1177/0961000620921930>. Acesso em: 24 abr. 2022.

SILVA, Antonio Braz de Oliveira *et al.* Análise de redes sociais como metodologia de apoio para a discussão da interdisciplinaridade na ciência da informação. **Ciência da Informação**, v. 35, p. 72-93, 2006.

SILVA, Jonathas Luiz Carvalho. **A identidade da Ciência da Informação brasileira no contexto das perspectivas históricas da Pós-Graduação**: análise dos conteúdos programáticos dos PPGCI'S. 2011. 229f. Dissertação (Mestrado em Ciência da Informação) – Centro de Ciências Sociais e Aplicadas, Universidade Federal da Paraíba, João Pessoa, 2011. Disponível em: <https://repositorio.ufpb.br/jspui/bitstream/tede/3905/1/arquivototal.pdf>. Acesso em: 24 abr. 2022.

SILVA, Tiago José da. **Historiografia e epistemologia no campo da ciência da informação**: um olhar sobre a literatura brasileira. 2019. 207f. Tese (Doutorado em Ciência da Informação) – Centro de Ciências Sociais e Aplicadas, Universidade Federal da Paraíba, João Pessoa, 2019. Disponível em: https://repositorio.ufpb.br/jspui/bitstream/123456789/18843/1/TiagoJos%c3%a9DaSilva_Tese.pdf. Acesso em: 24 abr. 2022.

SILVA, Hemerson Soares da; LAZZARIN, Fabiana Aparecida. Preprints na comunicação científica: proposta de modelo para aceleração do sistema de comunicação científica. **Informação e Profissões**, Londrina, v. 8, n. 2, p. 150-170, jul./dez. 2019. Disponível em: <https://www.uel.br/revistas/uel/index.php/infoprof/article/view/37742/pdf#>. Acesso em: 20 abr. 2022.

SILVA, Camila Ferreira; LOPES, Rodrigo de Macedo. A comunidade científica de Alagoas: um olhar a partir dos seus grupos de pesquisa. **Pro-Posições**, Campinas, v. 31, n. 1, p. 1-22, jan. 2020. Disponível em: <https://www.scielo.br/j/pp/a/QSMSPnfrh9cskpQ6zCvnmnp/?format=pdf&lang=pt>. Acesso em: 20 abr. 2022.

SILVA, Jorge Raimundo da; TARGINO, Maria das Graças. Visibilidade e prestígio na construção da rede colaborativa dos docentes de Medicina Veterinária da Universidade Federal de Alagoas: um olhar a partir do conceito de capital simbólico de Pierre Bourdieu. **Ciência da Informação em Revista**, Alagoas, v. 5, n. 2, p. 14-30, jul./dez. 2018. Disponível em: <https://www.seer.ufal.br/index.php/cir/article/view/4657>. Acesso em: 20 abr. 2022.

SILVA, Márcia Regina *et al.* Índice de autores normalizados e o processo de recuperação da informação: implicações para construção de indicadores bibliométricos. In: HAYASHI, Carlos Roberto Massão *et al.* (org.). **Bibliometria e Cientometria**: estudos temáticos. São Carlos: Pedro & João, 2013. p. 257-270.

SILVA, Francisco Edson Rodrigues da *et al.* Gestão de eventos: um estudo bibliométrico e sociométrico da produção científica internacional. **Revista de Gestão e Secretariado**, São Paulo, v. 12, n. 2, p. 122-146, maio/ago. 2021. Disponível em: <https://www.revistagesec.org.br/secretariado/article/view/1177>. Acesso em: 26 abr. 2022.

SILVA, Edson Marchetti da; SOUZA, Renato Rocha. Fundamentos em processamento de linguagem natural: uma proposta para extração de bigramas. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, v. 19, n. 40, p. 1-31, 2014.

SILVA, Verneck Ferreira da; KRONEMBERGER, Carlos Francisco; GOMES, Josir Simeone. Estudo bibliométrico: controle gerencial na internacionalização de empresas na base de dados da BDTD. **Brazilian Journal of Development**, Curitiba, v. 7, n. 10, p. 100147-100163, out. 2021. Disponível em: <https://www.brazilianjournals.com/index.php/BRJD/article/view/38337>. Acesso em: 30 jul. 2022.

SILVA, Fernanda Ferreira da. **Redes de colaboração científica nos cursos de Graduação em Biblioteconomia no Estado de São Paulo**: interseções entre Graduação e Pós-Graduação. 2018. 130f. Dissertação (Mestrado) – Escola de Comunicação e Artes, Universidade de São Paulo, São Paulo, 2018. Disponível em: <http://eprints.rcelis.org/38509/1/FernandaFerreiradaSilvaVC.pdf>. Acesso em: 26 abr. 2022.

SILVA, João Henrique da; PETERNELLA, Alessandra. Currículo e Educação Especial no/do campo: O que dizem as pesquisas sobre o tema. **Revista Brasileira de Educação do Campo**, v. 5, p. e8967-e8967, 2020.

SILVA, Armando Malheiro da; RIBEIRO, Fernanda. Ciência da informação trans e interdisciplinar: para a superação de equívocos. **Ciência da Informação: visões e tendências**, 2020.

SILVA, Márcia Danielly Cavalcanti; FIÚZA, Ana Louise De Carvalho; PINTO, Neide Maria de Almeida. The paradigmatic field of usage of the theoretical category of pluriactivity in Brazil. **Ciência Rural**, v. 49, n. 5, p. e20180660, 2019.

SILVA, Sônica Mônica da. **A produção científica da Ciência da Informação no Brasil**: análise do período de 2010 a 2020 nas bases de dados Web of Science e Scopus. 2021. 106f. Dissertação (Mestrado em Gestão e Organização do Conhecimento) – Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2021. Disponível em: <https://l1nq.com/XILRB>. Acesso em: 1 mar. 2022.

SILVA, Sônica Mônica da; OLIVEIRA, Marlene. A produção científica da Ciência da Informação no Brasil: análise do período de 2010 a 2020 nas bases de dados Web of Science e Scopus. **Pesquisa Brasileira em Ciência da Informação e Biblioteconomia**, João Pessoa, v. 16, n. 1, p. 1-14, jan./jun. 2021. Disponível em: <https://periodicos.ufpb.br/ojs/index.php/pbcib/article/view/58032/33089>. Acesso em: 1 mar. 2022.

SIMMHAN, Yogesh L.; PLALE, Beth; GANNON, Dennis. A survey of data provenance in e-science. **ACM Sigmod Record**, v. 34, n. 3, p. 31-36, 2005.

SIMÕES, Maria das Graças de Melo. Resumo documental: uma incursão à (des)construção concetual na literatura científica. Coimbra: Imprensa da Universidade de Coimbra, 2015. Disponível em: <https://doi.org/http://dx.doi.org/10.14195/978-989-26-1035-1>. Acesso em: 11 maio 2022.

SOARES, Filipi Miranda; MACULAN, Benildes Coura Moreira dos Santos; LIMA, Gercina Ângela Borém. Grupo de Pesquisa MHTX: dados bibliométricos da produção científica. **Revista ACB: Biblioteconomia em Santa Catarina**, v. 20, n. 3, p. 592-605, 2015.

SOBRAL, Natanael Vitor *et al.* Redes de colaboração científica na produção de conhecimento em doenças tropicais negligenciadas no Brasil: estudo a partir da Plataforma Lattes do CNPq. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, Florianópolis, v. 25, n. 1, p. 1-22, jan./jun. 2020. Disponível em: <https://brapci.inf.br/index.php/res/download/142248>. Acesso em: 26 abr. 2022.

SOUSA, Raquel Juliana Prado Leite de. Produção científica sobre letramento: mapeamento bibliométrico das teses da BDTD (1997-2016). **Cadernos de Pesquisa**, v. 50, n. 176, p. 494-514, 2020.

SOUSA, Evangelina da Silva; FONTENELE, Raimundo Eduardo Silveira. Mapeamento da produção científica internacional sobre Valores Humanos Básicos. **Em Questão**, Porto Alegre, v. 25, n. 3, p. 214-245, mar. 2019. Disponível em: <https://www.seer.ufrgs.br/EmQuestao/article/view/88249>. Acesso em: 30 jul. 2022.

SOUZA, Allan Rocha de *et al.* **Marcos legais nacionais em face da abertura de dados para pesquisa em saúde**: Dados pessoais, sensíveis ou sigilosos e propriedade intelectual. Rio de Janeiro: Fiocruz. 2018.

SOUZA, Francisco das Chagas Araújo *et al.* Perfil de pesquisadores científicos das regiões nordeste e sudeste do Brasil. **Research, Society and Development**, Marília, v. 11, n. 3, p. 1-7, mar. 2022. Disponível em: <https://rsdjournal.org/index.php/rsd/article/view/26334>. Acesso em: 20 abr. 2022.

SOUZA, Maria Tereza Saraiva de *et al.* Estudo bibliométrico de teses e dissertações em administração na dimensão ambiental da sustentabilidade. **REAd**, Porto Alegre, v. 76, n. 3, p. 541-568, set./dez. 2013. Disponível em: <https://www.scielo.br/j/read/a/HsH5n9KPKNcVb4763CgM3Sw/?format=pdf&lang=pt>. Acesso em: 30 jul. 2022.

SOUZA, Kelma Patrícia de; MARICATO, João de Melo. Motivações para a formação de redes de colaboração em bancas examinadoras de defesas de mestrado e doutorado. **InCID: Revista de Ciência da Informação e Documentação**, Ribeirão Preto, v. 12, n. 2, p. 173-193, fev. 2022. Disponível em: <https://www.revistas.usp.br/incid/article/view/181945/177598>. Acesso em: 16 fev. 2022.

SOUZA, Vanessa Faria de; PERRY, Gabriela Trindade. Mineração de Texto em Moocs: análise da relevância temática de postagens em fóruns de discussão. **Revista Novas Tecnologias na Educação**, Porto Alegre, v. 17, n. 13, p. 1-10, dez. 2019. Disponível em: <https://www.lume.ufrgs.br/bitstream/handle/10183/223653/001127865.pdf?sequence=1&isAllowed=y>. Acesso em: 26 abr. 2022.

SOUZA, Rosali Fernandez de; STUMPF, Ida Regina. Ciência da Informação como área do conhecimento: abordagem no contexto da pesquisa e da Pós-Graduação no Brasil. **Perspectiva em Ciência da Informação**, Belo Horizonte, v. 14, n. esp., p. 41-58, jul./dez. 2009. Disponível em:

<https://www.scielo.br/j/pci/a/bqYG8Pmk9kpTnDQPnCWjMPM/?format=pdf&lang=pt>.

Acesso em: 26 maio 2022.

SOUZA, Luana Quintal de. **As teses e dissertações do Programa de Pós-Graduação em Ciência da Informação da UFF: um estudo bibliométrico**. 2021. 180f. Dissertação (Mestrado em Ciência da Informação) – Instituto de Arte e Comunicação Social, Universidade Federal Fluminense, Niterói-RJ, 2021. Disponível em: <https://app.uff.br/riuff/handle/1/23206>. Acesso em: 25 abr. 2022.

STERNE, Jonathan AC *et al.* Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. **Bmj**, v. 338, 2009.

SUN, Jie; YUAN, Bao. Bibliometric mapping of the main articles in Librarianship and Information Science based on the Essential Science Indicators Database. **Journal of the Malaysian Library and Information Science**, Malásia, v. 25, n. 2, p. 61-76, jul./dez. 2020. Disponível em: <http://ijie.um.edu.my/index.php/MJLIS/article/view/26067>. Acesso em: 26 abr. 2022.

STUEBER, Ketlen; TEIXEIRA, Maria do Rocio Fontoura. Redes de conhecimento na comunicação científica em âmbito formal: panorama via Biblioteca Brasileira de Teses e Dissertações (2008-2018). **AtoZ**, v. 8, n. 2, p. 91-979, jul./dez. 2019.

STODDER, David; MATTERS, Why Data Preparation. Improving data preparation for business analytics. **Transforming Data With Intelligence**, v. 1, n. 1, p. 41, 2016.

TELMO, Flávia de Araújo. **Análise de redes sociais de colaboração em bancas de defesa de doutorado na pós-graduação em ciência da informação**. 2019. 160f. Dissertação (Mestrado em Ciência da Informação) – Centro de Ciências Sociais e Aplicadas, Universidade Federal da Paraíba, João Pessoa, 2019. Disponível em: <https://repositorio.ufpb.br/jspui/bitstream/123456789/17150/1/Arquivototal.pdf>. Acesso em: 15 fevereiro 2022.

THORPE, Geoffrey L.; FAVIA, Andrej. Data Analysis Using Item Response Theory Methodology: An Introduction to Selected Programs and Applications. **Psychology Faculty Scholarship**. 20. Disponível em: https://digitalcommons.library.umaine.edu/psy_facpub/20

TORINO, Emanuelle.; TREVISAN, Gustavo Lunardelli.; VIDOTTI, Silvana Aparecida Borsetti Gregorio. Dados abertos CAPES: um olhar à luz dos desafios para publicação de dados na web. **Ciência da Informação**, Brasília, v. 48, n. 3, p. 38-46, 2019. Disponível em: <http://repositorio.utfpr.edu.br/jspui/bitstream/1/4812/1/dadosabertosdesafiosdisponibilizacaoweb.pdf>. Acesso em: 10 jan. 2020.

TRINDADE, Eneus; CARVALHO, Priscila Rezende; PALETTA, Francisco Carlos. Estudo bibliométrico sobre moda em teses e dissertações na área das Ciências da Comunicação. **Em Questão**, v. 27, n. 3, p. 251-274, 2021.

TUESTA, Esteban Fernandez *et al.* Análise da participação das mulheres na ciência: um estudo de caso da área de Ciências Exatas e da Terra no Brasil. **Em Questão**, Porto Alegre, v. 25, n. 1, p. 37-62, jan./abr. 2019. Disponível em:

<https://www.seer.ufrgs.br/index.php/EmQuestao/article/view/80193/50605>. Acesso em: 30 jul. 2022.

TUKEY, John W. *et al.* **Exploratory data analysis**. Boston, EUA: Addison-Wesley Publishing Company, 1977.

UBALDI, Barbara. Open Government Data: towards empirical analysis of Open Government Data Initiatives. **OECD Working Papers on Public Governance**, n. 22, p. 1-61, 2013. Disponível em: <https://encr.pw/ZuHHr>.

UEHARA, Flavia Maria *et al.* O uso de objetos de aprendizagem no ensino fundamental: um estudo bibliométrico. **Revista Ibero-Americana de Estudos em Educação**, p. 1539-1553, 2017. Disponível em: <https://periodicos.fclar.unesp.br/iberoamericana/article/view/10308>

UNIVERSIDADE DE SÃO PAULO (USP). **Mestrado Profissional em Gestão da Informação**. São Paulo, 2022. Disponível em: <https://www3.eca.usp.br/pos/ppgci/mestrado-profissional>. Acesso em: 27 maio 2022.

UNIVERSIDADE DO ESTADO DE SANTA CATARINA (UDESC). Programa de Pós-Graduação em Gestão da Informação. **Apresentação**. Florianópolis, 2022. Disponível em: <https://www.udesc.br/faed/ppginfo/ppginfo>. Acesso em: 27 maio 2022.

UNIVERSIDADE DO ESTADO DE SANTA CATARINA (UDESC). Programa de Pós-Graduação em Gestão da Informação. **Apresentação**. Florianópolis, 2022. Disponível em: <https://www.udesc.br/faed/ppginfo/ppginfo>. Acesso em: 27 maio 2022.

UNIVERSIDADE FEDERAL DE SERGIPE (UFS). Programa de Pós-Graduação em Ciência da Informação. **Apresentação**. Aracaju, 2022. Disponível em: https://www.sigaa.ufs.br/sigaa/public/programa/apresentacao.jsf?lc=pt_BR&id=1051. Acesso em: 27 maio 2022.

UNIVERSIDADE FEDERAL DO CARIRI (UFCA). Programa de Pós-Graduação em Biblioteconomia. **Apresentação**. Cariri, 2022. Disponível em: <https://ppgb.ufca.edu.br/>. Acesso em: 27 maio 2022.

UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO (UNIRIO). Programa de Pós-Graduação em Biblioteconomia. **Bem-vindo ao PPGb-UNIRIO**. Rio de Janeiro, 2022a. Disponível em: <http://www.unirio.br/ppgb>. Acesso em: 27 maio 2022.

UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO (UNIRIO). **Programa de Pós-Graduação em Gestão de Documentos e Arquivos**. Rio de Janeiro, 2022b. Disponível em: <http://www.unirio.br/ppgarq>. Acesso em: 27 maio 2022.

UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE (UFRN). Programa de Pós-Graduação em Gestão da Informação e do Conhecimento. **Apresentação**. Natal, 2022. Disponível em: https://sigaa.ufrn.br/sigaa/public/programa/apresentacao.jsf?lc=pt_BR&id=9196. Acesso em: 27 maio 2022.

VANROSSUM, Guido; DRAKE, Fred. **The python language reference**. Amsterdam, Netherlands: Python *Software* Foundation, 2010.

VANZ, Sa. de S. A produção discente em comunicação no Brasil: análise das citações das dissertações defendidas no PPGCOM-UFRGS. In: XXV CONGRESSO BRASILEIRO DE CIÊNCIA DA COMUNICAÇÃO, 1 a 5 de setembro de 2002, Salvador. **Anais [...]**. Salvador: Sociedade Brasileira de Estudos Interdisciplinares da Comunicação, 2002.

VANTI, Nadia Aurora Peres. Da bibliometria à webometria: uma exploração conceitual dos mecanismos utilizados para medir o registro da informação e a difusão do conhecimento. **Ciência da informação**, v. 31, p. 369-379, 2002. Disponível em: <https://www.portcom.intercom.org.br/pdfs/117504444362926915389649162958472349166.pdf>

VERBORGH, Ruben; DE WILDE, Max. **Using OpenRefine**. [S. l.]: Packt Publishing Ltd, 2013.

VIANA, Andreia Soares. **Temáticas das teses dos Programas de Pós-Graduação em Ciência da Informação nível seis na Capes**. 2016. 154f. Dissertação (Mestrado em Ciência da Informação) – Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2016. Disponível em: https://repositorio.ufmg.br/bitstream/1843/BUBD-AE3JQF/1/disserta_o_andreia_cd.pdf. Acesso em: 3 fev. 2022.

VIANA, Andreia Soares; OLIVEIRA, Marlene. Teses dos programas de pós-graduação em ciência da informação nível seis na CAPES em relação às linhas de pesquisa. **Biblionline**, v. 12, n. 2, p. 117-131. 2016.

VIEIRA, Leticia Becker *et al.* Social phenomenology: analysis of academic production of graduate in Nursing, Brazil, 1979-2010. **Revista de Pesquisa Cuidado é Fundamental Online**, v. 5, n. 4, p. 652-660, 2013.

WERSIG, Gernot. Information science: the study of postmodern knowledge usage. **Information processing & management**, v. 29, n. 2, p. 229-239, 1993. Disponível em: <https://www.emerald.com/insight/content/doi/10.1108/00220411011087832/full/html>. Acesso em: 12 abr. 2025.

WERSIG, Gernot; NEVELING, Ulrich. The phenomena of interest to information science. **The information scientist**, v. 9, n. 4, p. 127-140, 1975.

WICKHAM, Hadley; GROLEMUND, Garrett. **R for data science**: Visualize, model, transform, tidy, and import data. Verlag nicht ermittelbar, 2017.

WILKINSON, Mark D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. **Scientific Data**, v. 3, n. 160018, p. 1-9, 2016. Disponível em: <https://www.nature.com/articles/sdata201618>

WITTEN, Ian H. *et al.* Practical machine learning tools and techniques. In: **Data Mining**. Amsterdam, The Netherlands: Elsevier, 2005. p. 403-413.

WOS. **Sobre a base**. [S. l.], 2022. Disponível em: <https://www-webofscience.ez27.periodicos.capes.gov.br/wos/history>. Acesso em: 18 jun. 2022.

ZHANG, Li; EICHMANN, Nickoal. Mapping the Scholarly Literature Found in Scopus on “Research Data Management”: a Bibliometric and Data Visualization Approach. **Journal of Librarianship and Scholarly Communication**, [S. l.], v. 7, n. 1, p. 1-20, jan./jun. 2019. Disponível em: <https://www.iastatedigitalpress.com/jlsc/article/12841/galley/12521/view/>. Acesso em: 26 abr. 2022.

ZUPIC, Ivan; ČATER, Tomaž. Bibliometric methods in management and organization. **Organizational research methods**, v. 18, n. 3, p. 429-472, 2015.

APÊNDICE A – SÍNTESE DOS ESTUDOS DA QUARTA ETAPA DA RIL

Este apêndice apresenta a planilha analítica que sistematiza os 29 estudos selecionados na revisão integrativa da literatura, conforme descrito na seção 3.3 da presente tese. A planilha foi estruturada com base em sete categorias de análise: título, autores, objetivo e conclusão, método aplicado (incluindo técnicas bibliométricas e softwares utilizados), visualização de dados, redes de colaboração e ano de publicação.

A organização cronológica dos estudos permite visualizar a evolução temporal das pesquisas, enquanto a categorização revela tendências metodológicas, temáticas recorrentes e lacunas que ainda demandam investigação. Este material serve de suporte empírico para as análises desenvolvidas na etapa 3.4 e contribui para o posicionamento da presente pesquisa frente ao estado da arte.

Quadro 106 – Síntese dos estudos constantes da amostra para análise

Título/temática	Autores	Objetivo e conclusão	Método utilizado (técnicas bibliométricas e <i>softwares</i>)	Visualização de dados	Redes de colaboração	Ano	Base
Análise de teses e dissertações em Avaliação Psicológica disponíveis na BVS-PSI Brasil	Maria Cristina Rodrigues Azevedo Joly <i>et al.</i>	Verificar a produção científica de teses e dissertações em avaliação psicológica no Brasil. Concluiu-se que os construtos mais estudados são a personalidade e a inteligência, e as áreas de aplicação em que se utiliza mais avaliação psicológica são a clínica e a escolar, de acordo com os resultados do presente estudo.	Coleta bibliométrica e manual nas teses e dissertações disponíveis na Biblioteca Virtual em Saúde do Brasil (Bvs-Psi Brasil).	Tabelas	Não realizado no estudo.	2010	LISA
A bibliometria na pesquisa científica da Pós-Graduação brasileira de 1987 a 2007	Ronaldo Araújo e Lúcia Alvarenga	Verificar a inserção dos estudos Bibliométricos na pesquisa científica da Pós-Graduação no Brasil. Concluiu-se acerca do interesse pela abordagem bibliométrica nas várias áreas de conhecimento no Brasil.	Levantamento no banco de dados da CAPES, com uso do <i>software Excel</i> para organização dos dados em uma planilha.	Gráficos, tabelas e quadros gerados de modo manual e com uso do <i>Excel</i> .	Não realizado no estudo.	2011	BRAPCI

Título/temática	Autores	Objetivo e conclusão	Método utilizado (técnicas bibliométricas e <i>softwares</i>)	Visualização de dados	Redes de colaboração	Ano	Base
Transgênicos e sociedade na agenda das pesquisas acadêmicas brasileiras	Maria Cristina Hayashi, Camila Rigolin e Carlos Hayashi	Situar o estado do conhecimento sobre a temática transgênicos e de seus impactos. Concluiu-se que os indicadores de produção científica sobre a temática revelaram as marcantes presenças em diversas áreas do conhecimento.	Estudo bibliométrico da produção científica acadêmica depositada no Banco de Teses da CAPES/MEC com uso do <i>Excel</i> para organização dos dados e do <i>software Vantage Point</i> para análise dos dados.	Tabelas e figuras	Não realizado no estudo.	2012	LISA
Teses e dissertações de enfermeiros sobre educação em enfermagem e saúde: um estudo bibliométrico	Vânia Marli Backes <i>et al.</i>	Descrever o panorama da produção científica em Educação na Enfermagem e na Saúde. Concluiu-se que a representatividade da temática no contexto nacional é significativa e contribui para o avanço do conhecimento na área da Enfermagem.	Coleta de dados nos catálogos de teses e dissertações do CEPEn/ABE, por meio do gerenciador EndNote e planilha Word.	Gráficos e tabelas com uso do <i>Excel</i> e Word.	Não realizado no estudo.	2013	LISA e SCOPUS
Social phenomenology: analysis of academic production of graduate in nursing, Brazil, 1979-2010/fenomenologia social	Letícia Vieira <i>et al.</i>	Analisar a produção acadêmica de Enfermagem brasileira com abordagem da Fenomenologia Social de Alfred Schütz. Concluiu-se que o referencial constitui-se como uma possibilidade na construção do conhecimento resultante de pesquisa em Enfermagem.	Coleta manual nos catálogos do Centro de Estudos e Pesquisas em Enfermagem de dissertações e teses que foram produzidas no período de 1979/2010.	Gráficos e tabelas.	Não realizado no estudo.	2013	LISA
Produção de teses e dissertações de Enfermagem em Portugal, 2000-2010: um estudo bibliométrico	Maria Aparecida Baggio <i>et al.</i>	Identificar as pesquisas da Enfermagem portuguesa, produzidas no período de 2000 a 2010. Concluiu-se que os enfermeiros pesquisadores, ao terem conhecimento das especificidades dos estudos, podem aperfeiçoá-las para a qualidade e excelência da formação de mestres e doutores.	Levantamento das dissertações da Universidade Católica Portuguesa e do Instituto de Ciências Biomédicas da Universidade do Porto e das teses de doutorado dessas instituições e da Universidade de Lisboa. Uso do <i>software Excel</i> .	Gráfico e tabelas.	Não realizado no estudo.	2014	SCOPUS
Grupo de pesquisa MHTX: dados	Filipi Soares, Benildes Maculan e	Identificar autores mais citados do MHTX. Concluiu-se que os integrantes	Mapeamento manual das características principais dos	Gráficos gerados pelo <i>Excel</i> .	Não realizado.	2015	BRAPCI

Título/temática	Autores	Objetivo e conclusão	Método utilizado (técnicas bibliométricas e <i>softwares</i>)	Visualização de dados	Redes de colaboração	Ano	Base
bibliométricos da produção científica	Gercina Lima	do grupo seguem uma mesma base teórico-metodológica, desde a tese que deu origem às pesquisas subsequentes.	estudos oriundos das teses e dissertações.				
Teses e dissertações em Enfermagem baseadas na Teoria das Representações Sociais: análise bibliométrica	Kalyane Oliveira, Clécio Maia e Francisco Miranda	Descrever o panorama da produção científica em Enfermagem baseada na Teoria das Representações Sociais. Concluiu-se que a representatividade do tema no contexto brasileiro contribui para o avanço do conhecimento na área da Enfermagem.	Coleta realizada no banco de dados da BDTD, com uso do <i>Excel</i> .	Tabelas e gráficos.	Não realizado.	2016	LISA
Gênero e sexualidade como temas de teses e dissertações: levantamento quantitativo nos repositórios do IBICT e da CAPES	Francisco Rios e Maria Naires Alves de Souza	Apresentar os registros numéricos da produção acadêmica de Pós-Graduação <i>stricto sensu</i> que abordam os temas gênero e sexualidade. Concluiu-se acerca da existência de uma quantidade tímida de estudos, tendo em vista a totalidade dos dados numéricos apresentados.	Coleta manual nos repositórios de teses e dissertações da CAPES e IBICT, quantificando a produção sobre a temática.	Tabelas com dados preenchidos manualmente.	Não realizado.	2017	BRAPCI
Panorama bibliométrico das teses e dissertações sobre educação indígena	Alexandre Masson Maroldi <i>et al.</i>	Elaborar um panorama bibliométrico da produção científica sobre educação indígena. Concluiu-se que as regiões Sudeste e Sul, respectivamente, concentram o maior número de trabalhos e que a área da Educação é majoritária na produção científica.	Coleta manual no banco de dados da BDTD/IBICT, com apresentação dos dados: programas, instituições, datas, orientadores, discentes, dentre outros.	Tabelas, quadros e gráficos.	Não realizado.	2017	LISA
O uso de Objetos de Aprendizagem no Ensino Fundamental: um estudo bibliométrico	Flávia Maria Uehara <i>et al.</i>	Investigar como a utilização de Objetos de Aprendizagem no contexto escolar, com alunos do Ensino Fundamental, vem ocorrendo e como está sendo discutida no país. Concluiu-se que o interesse pela temática tem aumentado nos últimos anos, nacionalmente, uma vez que ele é discutido por pesquisadores de vários estados e	Coleta bibliométrica manual no banco de dados de teses e dissertações da CAPES.	Tabelas e quadros.	Não realizado.	2017	LISA

Título/temática	Autores	Objetivo e conclusão	Método utilizado (técnicas bibliométricas e <i>softwares</i>)	Visualização de dados	Redes de colaboração	Ano	Base
		regiões do país.					
Scientific production of the International Classification for Nursing Practice: a bibliometric study.	Patrícia Josefa Fernandes Beserra <i>et al.</i>	Descrever a produção científica da Classificação Internacional para a Prática de Enfermagem a partir das dissertações e teses publicadas por enfermeiros no Brasil. Concluiu-se que o cenário da produção científica brasileira sobre a CIPE® é expressivo, evidenciando esse sistema como instrumento que possibilita a prestação de cuidados sistematizados.	Coleta bibliométrica manual realizada nos sites do Centro de Estudos e Pesquisas em Enfermagem, nos bancos de teses e dissertações da CAPES e da Plataforma Sucupira.	Gráficos e tabelas.	Não realizado.	2018	LISA
The use of the Calgary model in nursing dissertations and theses: a bibliometric study.	Solange Fátima Geraldo da Costa <i>et al.</i>	Analisar indicadores bibliométricos de dissertações e teses de Enfermagem sobre o Modelo Calgary de Avaliação Familiar. Concluiu-se que o modelo despertou o interesse dos pesquisadores em disseminarem o conhecimento produzido por meio de prestigiosos periódicos científicos da área.	Coleta de dados nos portais: Portal Capes, BDENF, BDTD e ao CEPEn. Uso do Programa <i>Excel</i> 14.0 e do Programa CmapTools, para elaborar mapa conceitual.	Tabelas, figuras e mapa conceitual.	Não realizado.	2018	LISA
Protestantismo, história e cultura escrita: um balanço da produção discente na Pós-Graduação brasileira	Jaime Magalhães Sepulcro Júnior e Ana Maria de Oliveira Galvão	Apresentar os resultados de pesquisa que consistiu na análise da relação entre protestantismo e cultura escrita na produção discente da Pós-Graduação brasileira.	Mapeamento na plataforma da BDTD do IBICT.	Gráficos	Não realizado.	2018	WOS
Dissertações e teses da Enfermagem acerca do Serviço de Atendimento Móvel de Urgência: estudo bibliométrico	Maria Aparecida Costa Agra <i>et al.</i>	Analisar a produção brasileira de dissertações e teses em Enfermagem que abordem o Serviço de Atendimento Móvel de Urgência. Concluiu-se acerca de uma limitada quantidade de trabalhos sobre o Serviço de Atendimento Móvel de Urgência nos Programas de Pós-Graduação em Enfermagem e lacunas relativas à temática nas pesquisas	Estudo bibliométrico realizado por meio do acesso virtual ao Centro de Estudos e Pesquisa em Enfermagem e ao Portal de Teses e Dissertações da CAPES. Uso do <i>Excel</i> .	Gráfico e tabelas	Não realizado.	2018	SCOPUS

Título/temática	Autores	Objetivo e conclusão	Método utilizado (técnicas bibliométricas e <i>softwares</i>)	Visualização de dados	Redes de colaboração	Ano	Base
		desses Programas.					
Análise da produção acadêmica sobre a Ludicidade na prática docente: o Ensino Fundamental como etapa emergente	Marcio Antonio Raiol Santos e Humberto Pereira	Analisar a produção acadêmica sobre a Ludicidade na prática docente da Educação Básica (2013 a 2016), a fim de saber de que forma se caracteriza o Ensino Fundamental, quais são os fenômenos investigados e suas orientações epistemológicas. Concluiu-se que o Ensino Fundamental assume uma condição privilegiada na produção de conhecimento sobre a Ludicidade na prática docente.	Mapeamento na BDTD/CAPES, com uso dos descritores: “ludicidade”, “jogo”, “brincadeira” e “brinquedo”.	Sem recursos visuais (apresentação textual).	Não realizado.	2019	WOS
Redes de conhecimento na comunicação científica em âmbito formal: panorama via Biblioteca Brasileira de Teses e Dissertações (2008-2018)	Ketlen Stueber e Maria do Rocio Fontoura Teixeira	Apresentar o estado da arte de pesquisas de Pós-Graduação sobre redes de conhecimento na produção da comunicação científica nos últimos dez anos (2008-2018). Concluiu-se que e as pesquisas abordam a produção do conhecimento científico em diferentes objetos, seja na esfera de concepção com base em grupos de estudo e linhas de pesquisa, seja na esfera da documentação.	Busca simples pelo termo “redes de conhecimento” na BDTD do IBICT, cuja exploração foi conduzida pela análise de conteúdo.	Quadros e figuras	Foram encontradas sete redes de colaboração, envolvendo instituições de várias regiões brasileiras.	2019	WOS
Mapeamento da produção científica na BDTD do IBICT sobre a Pedagogia da Alternância de 2011 a 2018	Odaléia Alves da Costa e Anny Camila Lima Rodrigues	Construir o mapeamento da produção científica no período de 2011 a 2018 sobre Pedagogia da Alternância. Concluiu-se que a Pedagogia da Alternância vivenciada nos CEFFAs se apresenta como proposta articulada com a formação do sujeito crítico e participativo.	Coleta na BDTD do IBICT, com uso do <i>software</i> Iramuteq.	Quadros, figuras, nuvens de palavras e árvore de conectividade.	Foram mapeados estudos citados, de autoria individual ou coletiva.	2019	WOS
Perspectivas emergentes na pesquisa educacional	Vera Lucia Felicetti <i>et al.</i>	Analisar as perspectivas a partir das quais o campo educacional tem sido investigado e sua relação com	Mapeamento de teses e dissertações produzidas em Programas de Pós-Graduação em Educação entre	Quadros e gráficos	Não realizado.	2019	WOS

Título/temática	Autores	Objetivo e conclusão	Método utilizado (técnicas bibliométricas e <i>softwares</i>)	Visualização de dados	Redes de colaboração	Ano	Base
		características do paradigma emergente. Concluiu-se que há o interesse em analisar os fenômenos educacionais de forma holística.	2012 e 2016 da CAPES.				
The paradigmatic field of usage of the theoretical category of pluriactivity in Brazil	Márcia Danielly Cavalcanti Silva, Ana Louise de Carvalho Fiúza e Neide Maria de Almeida Pinto	Identificar as tendências interpretativas sobre “pluriatividade”, no Brasil. Concluiu-se que o objetivo investigado tem se mantido como um tema atual de pesquisa.	Levantamento de documentos na BDTD do IBICT, com uso do <i>software IRAMUTEQ</i> .	Gráfico e tabela.	Não realizado.	2019	WOS
Educação ambiental na formação docente inicial: um mapeamento das pesquisas brasileiras em teses e dissertações	Sirlene Donaiski Motin <i>et al.</i>	Analisar tendências de pesquisa sobre educação ambiental na formação de professores. Concluiu-se que, apesar do aumento significativo na abrangência dos cursos pesquisados em relação à educação ambiental, há desafios adequados às produções.	Levantamento na BDTD do IBICT, com organização dos dados em planilha <i>Excel</i> .	Tabelas.	Não realizado.	2019	SCOPUS
Dissertações e teses sobre Cuidados Paliativos em Oncologia Pediátrica: estudo bibliométrico	Kalina Coeli Costa de Oliveira Dias <i>et al.</i>	Analisar indicadores bibliométricos sobre Cuidados Paliativos em Oncologia Pediátrica defendidos em Programas de Pós-Graduação no cenário brasileiro. Como conclusão, a pesquisa identificou um número reduzido de estudos.	Estudo bibliométrico realizado, manualmente, nos bancos de dados da CAPES e da BDTD, com uso do <i>Excel</i> .	Tabelas e mapa conceitual.	Não realizado.	2020	LISA
Scientific production of dissertations and theses on Palliative Care and Chronic Diseases: bibliometric study	Thalis Regina Silva Paiva <i>et al.</i>	Verificar indicadores bibliométricos de dissertações e de teses publicadas no Brasil sobre Cuidados Paliativos e Doenças Crônicas. Concluiu-se que a maioria dos estudos se desenvolveu por meio da abordagem qualitativa e o cenário com maior quantitativo de pesquisa foi o local hospitalar.	Coleta bibliométrica nos bancos de dados da CAPES, da BDTD e do Portal Domínio Público. Uso do <i>software Excel</i> .	Figuras, tabelas e mapa conceitual.	Não realizado.	2020	LISA e WOS
Palliative Care in Primary Health Care: scientific	Leonia Maria de Lima <i>et al.</i>	Analisar a produção científica de Programas de Pós-Graduação <i>stricto</i>	Revisão bibliométrica realizada no Portal Domínio Público, BDTD,	Figuras, tabelas e mapa conceitual.	Não realizado.	2020	WOS

Título/temática	Autores	Objetivo e conclusão	Método utilizado (técnicas bibliométricas e <i>softwares</i>)	Visualização de dados	Redes de colaboração	Ano	Base
production of theses and dissertations in the Brazilian framework		<i>sensu</i> brasileiros sobre Cuidados Paliativos na Atenção Primária à Saúde. Concluiu-se que a produção científica de dissertações e teses produzidas pelos Programas de Pós-Graduação <i>stricto sensu</i> pode colaborar para a discussão entre gestores e profissionais de saúde.	Banco de Teses e Dissertações da USP, UFRJ, UFSC e Portal de Teses e Dissertações da CAPES. Uso do <i>Excel</i> , estatística descritiva e metodologia de mapeamento de conceitos.				
Taxonomia corporativa e taxonomia facetada: usos e aplicações na Ciência da Informação no Brasil	Pâmela Lopes, Elisângela Cristina Aganette e Benildes Maculan	Apresentar os usos e aplicações de taxonomias corporativas e facetadas como instrumentos de representação da informação e do conhecimento. Concluiu-se que o quantitativo pouco representativo de pesquisas evidencia a necessidade de mais estudos envolvendo esses instrumentos.	Levantamento bibliográfico na BDTD da CAPES, com uso da análise de conteúdo, mediante categorização.	Quadros	Não realizado.	2020	WOS
Currículo e Educação Especial no/do Campo: o que dizem as pesquisas sobre o tema?	João Henrique da Silva e Alessandra Peternella	Analisar a produção científica oriunda de teses e dissertações produzidas no Brasil sobre a temática da Educação Especial no/do Campo. Concluiu-se acerca da necessidade de estudos futuros e do avanço do tema.	Levantamento bibliométrico e análise de conteúdo nos documentos da BDTD do IBICT. Uso do <i>software</i> MS <i>Excel</i> .	Quadros, gráficos e tabelas.	Não realizado.	2020	WOS
Produção científica sobre Letramento: mapeamento bibliométrico das teses da BDTD (1997-2016)	Raquel Juliana Prado Leite de Sousa	Identificar a constituição do campo científico do Letramento no Brasil. Concluiu-se acerca da compreensão do campo científico sobre Letramento no Brasil pela identificação de algumas de suas raízes históricas.	Levantamento na BDTD, com uso dos <i>softwares</i> : <i>Excel</i> e <i>VantagePoint®</i> .	Gráficos e tabelas.	Não realizado.	2020	SCOPUS

Título/temática	Autores	Objetivo e conclusão	Método utilizado (técnicas bibliométricas e <i>softwares</i>)	Visualização de dados	Redes de colaboração	Ano	Base
Estudo bibliométrico sobre Moda em teses e dissertações na área das Ciências da Comunicação	Eneus Trindade, Priscila Rezende Carvalho e Francisco Carlos Paleta	Dimensionar e caracterizar a produção sobre Moda. Por meio do estudo, foi possível apresentar as instituições e os pesquisadores de destaque e relacionar as palavras-chave mais utilizadas, vislumbrando tendências de pesquisa.	Levantamento no banco de dados da CAPES, com uso dos <i>softwares</i> : <i>Excel</i> , <i>Iramuteq</i> e <i>RStudio</i> .	Figuras e gráficos elaborados com uso dos <i>softwares</i> utilizados no método.	Não realizado.	2021	BRAPCI e LISA
Teses de doutoramento em Gestão do Esporte no Brasil: uma revisão integrativa metodológica	Ivan Furegato Moraes, Cacilda Mendes dos Santos Amaral e Flávia da Cunha Bastos	Analisar temáticas e métodos sobre Gestão do Esporte. Concluiu-se que o panorama temático revela defasagens coerentes com as do desenvolvimento científico da área no Brasil.	Coleta manual no banco de dados da BDTD e descrição qualitativa por meio da análise de conteúdo (categorização e codificação).	Gráficos, tabelas e quadros.	Não realizado.	2021	LISA

Fonte: elaborado pelo autor (2022).