Produção Científica em Ciência da Informação: utilizando os dados abertos CAPES

Scientific Production in Information Science: using CAPES open data

Patrícia Ofélia Pereira de Almeida
Universidade Estadual de Londrina, Londrina-PR, pereira@uel.br

Patrick Stacy Meyer
Universidade Estadual de Londrina, Londrina-PR, patrick.enzo.meyer@gmail.com

Resumo:
A produção científica é um requisito fundamental para pesquisadores, e no âmbito da pós-graduação *stricto senso* faz parte dos requisitos necessários para a obtenção do título de mestre ou doutor. O presente estudo tem como objetivo descrever alguns aspectos dos metadados do Catálogo de Teses e Dissertações disponibilizados pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, no que se refere à sua estrutura, e aos processos de coleta, limpeza e utilização para fins de pesquisa e análise, e ainda apresentar um panorama da produção em Ciência da Informação. Tem características quantitativa, descritiva e analítica. Foram coletados os conjuntos de dados do Catálogo de Teses e Dissertações – Brasil, referentes aos anos de 2017 a 2020. Como resultados, observou-se as produções são mais recorrentes na região Sudeste, sendo que a maioria absoluta se refere a dissertações, e em seguida as teses. Houve uma queda no número de produções no ano de 2020.
Palavras-chave: Dados abertos; Ciência de dados; CAPES; Programas de pós-graduação em Ciência da Informação; Dissertações e teses.

Abstract:
Scientific production is a fundamental requirement for researchers, and within the stricto sensu graduate program it is part of the requirements for obtaining a master's or doctorate degree. The present study aims to describe some aspects of the metadata of the Theses and Dissertations Catalog made available by the Coordination for the Improvement of Higher Education Personnel, in terms of its structure, and the processes of collection, cleaning and use for research and analysis, and also to present an overview of the production in Information Science. It has quantitative, descriptive and analytical characteristics. Datasets from the Catalog of Theses and Dissertations - Brazil were collected, referring to the years 2017 to 2020. As a result, it was observed that the productions are more recurrent in the Southeast region, with the absolute majority referring to dissertations, and then the theses. There was a drop in the number of productions in the year 2020.
Keywords: Open data; Data Science; CAPES; Postgraduate programs in Information Science; Dissertations and theses.

1 Introdução

A produção científica é um requisito fundamental para pesquisadores, pois é uma forma de divulgarem os resultados parciais e finais de suas investigações científicas, e obter a avaliação e o reconhecimento de seus pares.

No âmbito da pós-graduação *stricto senso*, a publicação de artigos em periódicos, eventos, e outras formas de produção científica faz parte dos requisitos necessários para a obtenção do título de mestre ou doutor.

A Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) coleta e disponibiliza os dados referentes a produção científica proveniente da pós-graduação no Brasil, de forma que é possível obter, tratar e analisar tais dados sob diversas perspectivas.

O presente estudo tem como objetivo descrever alguns aspectos dos metadados do Catálogo de Teses e Dissertações disponibilizados pela CAPES, no que se refere à sua estrutura, e aos processos de coleta, limpeza e utilização para fins de pesquisa e análise, e ainda apresentar um panorama da produção em Ciência da Informação nos respectivos registros.

2 Referencial Teórico

Trabalhar com um elevado volume de dados tem sido uma preocupação constante, principalmente a partir da década de 60, quando as indústrias, as pesquisas, e a comunicação científica começaram a evoluir de forma mais ativa e com maior dimensão quantitativa de resultados.

A Ciência da Informação tem se preocupado com o tratamento de dados, no intuito de transformá-los em insumo de valor para a tomada de decisão, em especial depois que Borko (1968) definiu o escopo de estudo da Área.

Nesse sentido, Donoho (2015) afirma que a mais de 50 anos foi detectada a necessidade de trabalhar e aprender com dados, de estabelecer novos métodos de uso da estatística. Porém, continua o autor, faz somente cerca de uma década que as principais universidades têm investido nos programas de ciência de dados.

Para Grus (2021, p. 18), o cientista de dados é um profissional que detém conhecimentos acerca de estatística e computação, e utiliza suas habilidades para extrair conhecimento de dados desorganizados.

Nesse sentido, pode-se dizer que o *Data Science*, ou Ciência de Dados, consiste em utilizar recursos da computação para tratar/organizar uma grande quantidade de dados, aplicando modelos matemáticos e estatísticos, de forma que os resultados sejam sintetizados e a análise dos dados se torne possível. Nessa direção, Coneglian, Santarem Segundo e Sant'ana, (2017) consideram que com a análise de dados permite detectar padrões, que ao serem modelados se tornam informações que dão suporte ao processo de tomada de decisão.

A CAPES (2022) apresenta que o objetivo da avaliação da pós-graduação *stricto senso* no Brasil consiste na certificação da qualidade, assim como identificar discrepâncias regionais e de áreas estratégicas do conhecimento. Nesse sentido, identificar padrões de produção científica (ou a falta deles) no âmbito de um programa, Instituições de Ensino Superior (IES), regiões ou mesmo de forma global, é uma forma de diagnosticar problemas que podem estar prejudicando o rendimento dos programas, ou mesmo boas práticas que estejam elevando a produtividade.

Dessa forma, os dados disponibilizados pela CAPES são uma valiosa fonte de informação, que precisam ser tratados para que possam fornecer padrões e perspectivas acerca dos programas de pós-graduação *stricto sensu*, das áreas do conhecimento, e diversos outros aspectos possíveis.

Os metadados "crus" fornecidos pela CAPES são um apanhado de letras, números e símbolos sem sentido semântico, mas se tratados, constituem um rico estoque de informações que podem ser modeladas para finalidades específicas.

3 Procedimentos Metodológicos

O presente estudo tem características quantitativa, descritiva e analítica, visto que pretende apresentar e agrupar um volumoso conjunto de metadados.

Para atingir o objetivo proposto, foram coletados os conjuntos de dados do Catálogo de Teses e Dissertações – Brasil[1], referentes ao quadriênio de 2017 a 2020.

Foram capturados os arquivos em formato CSV, de forma que pudesse ser mais facilmente utilizado pelo software Microsoft Excel, considerando ser um recurso que habitualmente está disponível em uma grande quantidade de computadores pessoais, além de ser fácil encontrar na Internet tutoriais para usos de diversas aplicações. Obviamente existem outros *softwares* específicos para o tratamento de dados, que podem oferecer resultados mais rápidos e mais precisos, contudo a menor disponibilidade de tutoriais e a necessidade de treinamento do zero foram considerados como obstáculos, o que demanda de maior tempo de dedicação para ser superado.

Para a apresentação do panorama da produção em Ciência da Informação nos respectivos registros, foram selecionados os metadados categorizados com a respectiva área do conhecimento, com a finalidade de quantificação e apresentação dos resultados.

[1] https://dadosabertos.capes.gov.br/dataset/2017-2020-catalogo-de-teses-e-dissertacoes-da-capes.

4 Resultados

Os arquivos recuperados (.csv) apresentaram em média o tamanho de 445 MB e 87.599 registros. Apesar de serem um pouco pesados, a estrutura dos dados em planilhas se mostrou de fácil entendimento e modelagem.

A Tabela 1 apresenta o ano, a quantidade de IES, o número de programas de Pós-Graduação e o total de produções registradas em cada conjunto de dados coletados.

Tabela 1 – Dados da produção da pós-graduação no Brasil – 2017 a 2020.

ANO	IES	Programas	Produção	Cresc. anual (%)
2017	424	1.712	85.310	-
2018	441	1.788	90.469	6%
2019	453	1.828	94.503	4%
2020	466	1.819	80.114	-15%
Total			**350.396**	

Fonte: Dados da pesquisa.

Pode-se observar que o número geral de produções teve um crescimento gradativo de 2017 a 2019, mas regrediu bruscamente em 2020. É possível que essa queda se justifique pelo distanciamento social causado pela pandemia da Covid-19, cujos reflexos afetou todos os setores e, portanto, com a educação não foi diferente. O corte de bolsas e a impossibilidade de aulas presenciais, dentre outros aspectos, foram prejudiciais para a manutenção do calendário acadêmico. Apenas com os dados do quadriênio 2021/2024 será possível analisar os reflexos quantitativos que o período causou para a pós-graduação no Brasil.

Os dados disponibilizados pela CAPES estão estruturados em 58 campos, os quais podem ser códigos numéricos ou textuais. Não foi possível localizar uma tabela com as siglas utilizadas para identificação das colunas de dados, e em alguns casos foi necessário recorrer ao conteúdo do campo para identificar seu teor. O sistema categoriza os campos por prefixos (AN, CD, DH, DS, DT, ID, IN, NM, NR, SG).

Foi possível identificar que os campos apresentam dados dos documentos no que se refere à representação descritiva (autor, título, número de páginas etc.), temática (palavras-chave, resumo etc.), vínculo (Instituição, programa, projeto etc.), área (área do conhecimento, linha de pesquisa etc.), pessoal (categoria, titulação etc.) e de data.

Observou-se que os dados permitem uma variada gama de recortes para análises quantitativas ou qualitativas. Contudo, para o objetivo de demonstrar o panorama da produção em Ciência da Informação, optou-se por descartar alguns campos que não se mostraram representativos para esta finalidade, tornando o arquivo dos metadados mais leve e maleável. Nesse sentido, foram mantidos os dados básicos de identificação dos documentos, o que resultou em 17 campos.

Do total geral de 350.396 registros de documentos recuperados, foram selecionados os da Área do Conhecimento Ciência da Informação, e organizados em um novo arquivo. Os dados foram agrupados aplicando os recursos de tabela dinâmica, com o intuito de apresentar a representação visual da informação. Nas 21 IES identificadas, constam cadastradas um total de 1520 produções em CI (Tabela 2).

Tabela 2 – Produção da pós-graduação em Ciência da Informação no Brasil por instituição/ano – 2017 a 2020.

REGIÃO/ IES	2017	2018	2019	2020	Total
Centro-Oeste	**27**	**29**	**26**	**21**	**103**
UNB	27	29	26	21	103
Nordeste	**61**	**89**	**129**	**93**	**372**
FUFSE	-	-	15	18	33
UFBA	9	23	21	5	58
UFC		15	9	7	31
UFCA	3	13	16	18	50
UFPB-JP	25	12	39	25	101
UFPE	16	17	13	11	57
UFRN	8	9	16	9	42
Norte	-	**1**	**9**	**16**	**26**
UFPA	-	1	9	16	26
Sudeste	**163**	**232**	**228**	**185**	**808**
FCRB	-	11	13	11	35
FUMEC	-	28	19	24	71
UFF	9	10	23	15	57
UFMG	37	36	46	44	163
UFRJ	19	32	35	28	114
UFSCAR	-	9	10	10	29
UNESP-MAR	33	43	48	34	158
UNIRIO	49	27	20	9	105
USP	16	36	14	10	76
Sul	**59**	**47**	**53**	**52**	**211**
UDESC	13	15	13	12	53
UEL	19	11	10	13	53
UFSC	27	21	30	27	105

Total CI	**310**	**398**	**445**	**367**	**1.520**
% do total geral	**0,36**	**0,44**	**0,47**	**0,46**	**0,43**

Fonte: Dados da pesquisa.

Apesar da redução de produções registradas em 2020, a Tabela 2 demonstra que CI manteve um percentual anual de produções próximo da média dos anos anteriores e acima da média total, de forma que não se mostrou prejudicada em maior medida do que as outras áreas.

Também é possível observar que a UFMG e a UNESP-MAR são as instituições com maior número de produções no quadriênio (10,8% e 10,4%, respectivamente). Em seguida aparecem a UFRJ (7,5%), a UNIRIO e a UFSC (6,91%), a UNB (6,78%) e a UFPB-JP (6,64%). As demais instituições somaram 44,1%.

A maior produção em CI está na região Sudeste, que detém mais da metade do total (53,2%), onde também se concentra um maior número de instituições com programas de pós-graduação na Área. O Nordeste aparece em segundo lugar (24,5%), seguido do Sul (13,9%), Centro-Oeste (6,8%) e Norte (1,7%).

Pode-se observar também que a redução na produção foi linear em 2020, atingiu a maioria das instituições, apenas seis IES mantiveram ou aumentaram o número de defesas (FUFSE, UFCA, UFPA, FUMEC, UFSCAR e UEL).

A Tabela 3 apresenta a produção da pós-graduação em Ciência da Informação no Brasil por estado.

Tabela 3 – Produção da pós-graduação em Ciência da Informação no Brasil por estado – 2017 a 2020.

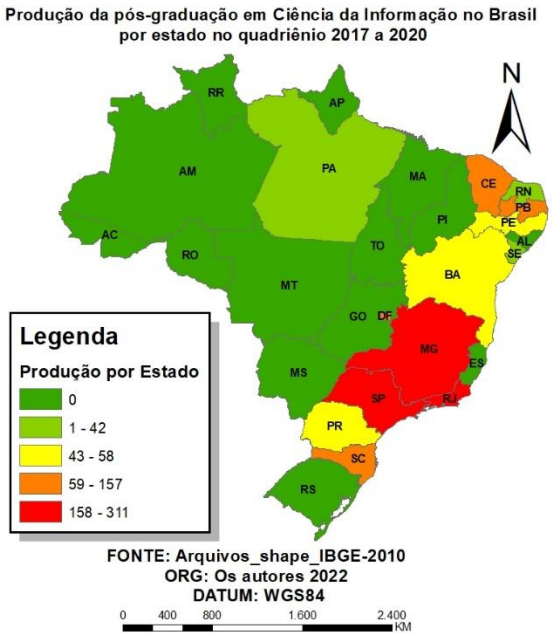
REGIÃO/UF	2017	2018	2019	2020	Total
Centro-Oeste	**27**	**29**	**26**	**21**	**103**
DF	27	29	26	21	103
Nordeste	**61**	**89**	**129**	**93**	**372**
BA	9	23	21	5	58
CE	3	28	25	25	81
PB	25	12	39	25	101
PE	16	17	13	11	57
RN	8	9	16	9	42
SE	-	-	15	18	33
Norte	-	**1**	**9**	**16**	**26**
PA	-	1	9	16	26
Sudeste	**163**	**232**	**228**	**185**	**808**
MG	37	64	65	68	234
RJ	77	80	91	63	311
SP	49	88	72	54	263
Sul	**59**	**47**	**53**	**52**	**211**
PR	19	11	10	13	53
SC	40	36	43	39	158
Total	**310**	**398**	**445**	**367**	**1.520**

Fonte: Dados da pesquisa.

Observa-se que o Sudeste é a região mais produtiva, na qual as instituições estão de certa forma equilibrada no número de produções registradas. Contudo, destaca-se o estado do Rio de Janeiro com um acumulado de produções mais significativo.

Esses dados podem ser visualizados na Figura 1, que destaca gradativamente a produção da pós-graduação *Stricto sensu* em CI por estado.

Figura 1 – Produção da pós-graduação em Ciência da Informação no Brasil por estado – 2017 a 2020.



Fonte: Dados da pesquisa.

A representação visual proporcionada pelo mapa permite visualizar com maior destaque onde estão as IES mais produtivas em CI.

Na Tabela 4 estão apresentados os dados da CI por tipo de produção.

Tabela 4 – Produção da pós-graduação em Ciência da Informação no Brasil por tipo – 2017 a 2020.

TIPO	2017	2018	2019	2020	Total
Dissertação	227	296	310	259	1.092
Tese	73	93	121	96	383
Produto, proced. ou técnica	10	9	14	4	37
Projeto técnico	-	-	-	6	6
Editoria	-	-	-	1	1

Relatório final de pesquisa	-	-	-	1	1
Total	**310**	**398**	**445**	**367**	**1.520**

Fonte: Dados da pesquisa.

Embora tenham sido identificadas mais de 20 tipos de produções no contexto geral dos dados coletados, na área de Ciência da Informação constatou-se a presença de somente seis tipos. As dissertações são significativamente a maioria absoluta das produções recuperadas. Em seguida figuram as teses que, embora estejam em menor número, ainda é mais representativo que os demais tipos de produção.

Ao todo, somaram-se 383 teses de doutorado, 802 dissertações de mestrado acadêmico, e no mestrado profissional são 290 dissertações e 45 produções de outros tipos. Identificou-se ainda 406 orientadores, contudo, apenas 62 (15,3%) estavam presentes nos quatro anos de dados coletados, e apenas sete (1,7%) orientaram 10 ou mais trabalhos defendidos. Cada orientador acompanhou de 1 a 15 pós-graduandos, perfazendo uma média de 3,8 defesas por orientador no período analisado.

5 Considerações Finais

Com base no exposto, foi possível utilizar os dados abertos da Capes para a realização da pesquisa, na qual identificou-se que houve uma queda da produção em 2020. Considerando que as dissertações lideram a maioria absoluta das produções, que a capacitação em nível de mestrado habitualmente ocorre no prazo de dois anos, associada à situação de pandemia que teve início no final do ano de 2019, pode-se supor que houve um atraso nas defesas, e que possivelmente o quantitativo que diminuiu em 2020 irá figurar no relatório CAPES de 2021.

Constatou-se também que a região Sudeste é um polo no que se refere à produção em CI, sendo que ali figuram também grandes centros de pesquisa e inovação tecnológica, mas que também sofreu uma queda em 2020.

Por fim, reitera-se a ideia de que a análise de dados permite uma série de perspectivas acerca daquilo que representam, e que os dados disponibilizados pela CAPES podem proporcionar inúmeras outras análises de cunho qualitativo.

6 Referências

BORKO, H. Information science: what is it? **American Documentation**, v. 19, n. 1, 1968.

CONEGLIAN, C. S.; SANTAREM SEGUNDO, J. E.; SANT'ANA, R. C. G. Big Data: fatores potencialmente discriminatórios em análise de dados. **Em Questão**, Porto Alegre, v. 23, n. 1, p. 62–86, 2017. Disponível em: https://seer.ufrgs.br/index.php/EmQuestao/article/view/62122. Acesso em: 2 set. 2022.

COORDENAÇÃO de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). **Dados abertos**. Brasília (DF): CAPES, 2022. Disponível em: https://dadosabertos.capes.gov.br. Acesso em: 05 set. 2022.

DONOHO, David. 50 Years of Data Science, **Journal of Computational and Graphical Statistics**, v. 26, n. 4, p. 745-766, 2017. Disponível em: https://www.tandfonline.com/doi/full/10.1080/10618600.2017.1384734. Acesso em: 17 set. 2022.

GRUS, Joel. **Data Science do zero**. 2. ed. Rio de Janeiro: Editora Alta Books, 2021.