



UNIMAR - UNESP/PPGCI - IBICT
MARÍLIA

VIII WORKSHOP DE INFORMAÇÃO,
DADOS E TECNOLOGIA

Unimar
UNIVERSIDADE DE MARÍLIA



WIDaT 2025

UNIMAR - UNESP/PPGCI - IBICT
MARÍLIA

ANAIS

VIII WORKSHOP DE INFORMAÇÃO,
DADOS E TECNOLOGIA

REALIZAÇÃO:

Unimar
UNIVERSIDADE DE MARÍLIA

unesp 

 **ibict**
Instituto Brasileiro de Informação
em Ciência e Tecnologia

COORDENAÇÃO GERAL

Caio Saraiva Coneglian (UNIMAR, Unesp)

VICE COORDENAÇÃO

Silvana Aparecida Borsetti Gregorio Vidotti
(Unesp, Ibict)



UNIMAR - UNESP/PPGCI - IBICT
MARÍLIA

COMISSÃO CIENTÍFICA

Emanuelle Torino (presidente da comissão) (UTFPR, Ibict)
Ana Carolina Simionato Arakaki (UnB)
Andre Luiz Apel (UNIR)
Caio Saraiva Coneglian (UNIMAR, Unesp)
Cecilio Merlotti Rodas (IFSP, Unesp)
Daniela Lucas da Silva Lemos (UFES)
Diana Vilas Boas Souto Aleixo (UFES)
Edberto Ferneda (Unesp)
Elvis Fusco (Fundação Nishimura)
Fabricio Silva Assumpção (UFSC)
Fernanda Alves Sanchez (Unesp)
Guilherme Ataíde Dias (UFPB)
Helen de Castro Silva Casarin (Unesp)
Henry Poncio Cruz de Oliveira (UFPB)
Jacquelin Teresa Camperos Reyes (UFPA)
Jean Fernandes Brito (Uniasselvi)
José Eduardo Santarém Segundo (USP, Unesp)
Lais Barbudo Carrasco (MIUN, Suécia)
Leonardo Castro Botega (UNIMAR, Unesp)
Luana Farias Sales Marques (Ibict)
Maíra Murrieta Costa (MCTI)
Mariana Baptista Brandt (Câmara dos Deputados)
Maria José Vicentini Jorente (Unesp)
Mariângela Spotti Lopes Fujita (Unesp)
Milton Shintaku (Ibict)
Natália Marinho do Nascimento (Unesp)
Rachel Cristina Vesu Alves (Unesp)
Rafael Gutierrez Castanha (UNIMAR)
Ricardo César Gonçalves Sant'Ana (Unesp)
Rosilene Paiva Marinho de Sousa (UFPB, Ibict)
Sandro Rautenberg (Unicentro)
Sandra de Albuquerque Siebra (UFPE)
Sandra Milena Roa-Martinez (UNICAUCA – Colômbia)
Silvana Aparecida Borsetti Gregorio Vidotti (Unesp, Ibict)
Walkiria Martinez Heinrich Ferrer (UNIMAR)

CAPA

Lucas Alves

DIAGRAMAÇÃO

Diego André Andrade Neves (UFPR)

NORMALIZAÇÃO

Clara Braz de Almeida (Ibict)
Joana Gentil Penna (Ibict)
Julia Rodrigues da Silva (Ibict)
Nicole Oliveira da Silva (Ibict)
Paloma Rayana França da Silva (Ibict)



4.0 Internacional

Esta licença permite compartilhamento, remixe, adaptação e criação a partir do trabalho, mesmo para fins comerciais, desde que sejam atribuídos créditos ao(s) autor(es). Conteúdos elaborados por terceiros, citados e referenciados nesta obra não são cobertos pela licença.

Desafios e avanços na implementação dos princípios de dados abertos no Brasil: uma análise dos planos de ação da INDA

Challenges and progress in implementing open data principles in Brazil: an analysis of INDA's action plans

Diana Vilas Boas Souto Aleixo¹

 <http://lattes.cnpq.br/5082397017911269>

 <https://orcid.org/0000-0003-4156-354X>

Ricardo Cesar Gonçalves Sant'Ana²

 <http://lattes.cnpq.br/1022660730972320>

 <https://orcid.org/0000-0003-1387-4519>

Resumo

O estudo realiza um diagnóstico sobre a implementação dos princípios de dados abertos nos Planos de Ação da Infraestrutura Nacional de Dados Abertos (2013-2014, 2017-2018, 2021-2022 e 2023-2025), com base na proposta de Aleixo (2020). A pesquisa caracteriza-se como aplicada, de natureza qualitativa, estabelecida como exploratória e descritiva, que fez uso da pesquisa bibliográfica e documental. Os resultados foram sistematizados por meio da análise documental categorizada, na qual os dados extraídos dos Planos de Ação da Infraestrutura Nacional de Dados Abertos foram organizados conforme os “8 Princípios de dados abertos” – Princípios de Sebastopol. A análise identificou avanços em acessibilidade e atualidade, mas destacou lacunas, especialmente na participação da sociedade civil e no reuso dos dados. Conclui-se que a ausência dos princípios nos documentos reguladores resulta em um cenário de anomia, enfraquecendo o potencial dos dados como ferramenta de controle social e inovação. É, portanto, essencial fortalecer as políticas públicas e promover uma colaboração mais eficaz entre o Estado e a sociedade para garantir a efetiva utilização dos dados abertos no Brasil.

Palavras-chave: anomia dos dados; dados governamentais abertos; planos de ação; princípios de dados abertos; Infraestrutura Nacional de Dados Abertos.

Abstract

The study carries out a diagnosis of the implementation of open data principles in the National Open Data Infrastructure Action Plans (2013-2014, 2017-2018, 2021-2022 and 2023-2025), based on Aleixo (2020). The research is characterized as applied, of a qualitative nature, established as exploratory and descriptive, which made use of bibliographic and documentary research. The results were systematized by means of categorized documentary analysis, in which the data extracted from the National Open Data Infrastructure Action Plans was organized according to the “8 Principles of Open Data” - Sebastopol Principles. The analysis

¹ Doutora em Ciência da Informação, Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual Paulista (PPGCI/UNESP), Marília, SP, Brasil. Docente, Departamento de Arquivologia, Universidade Federal do Espírito Santo (CCJE/UFES), Vitória, ES, Brasil. E-mail: diana.aleixo@ufes.br.

² Doutor em Ciência da Informação, Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual Paulista (PPGCI/UNESP), Marília, SP, Brasil. Docente, Universidade Estadual Paulista, Marília, SP, Brasil. E-mail: ricardo.santana@unesp.br.

identified advances in accessibility and timeliness, but highlighted gaps, especially in civil society participation and data reuse. It concludes that the absence of the principles in the regulatory documents results in a scenario of anomie, weakening the potential of data as a tool for social control and innovation. It is therefore essential to strengthen public policies and promote more effective collaboration between the state and society to ensure the effective use of open data in Brazil.

Keywords: data anomie; open government data; data plans; open data principles; National Open Data Infrastructure.

1 INTRODUÇÃO

Nas últimas décadas, observa-se que a abertura de dados governamentais tem se consolidado como um movimento global, impulsionado pelo avanço das tecnologias, e pela necessidade de transparência e participação social. Neste contexto, diversos países passaram a adotar legislações, infraestruturas e diretrizes técnicas baseadas em princípios internacionais, com o intuito de padronizar o acesso e promover o reuso dos dados públicos.

No Brasil, a adesão a iniciativa ocorreu em 2011, com o ingresso do país a *Open Government Partnership* (OGP)³, a ratificação da Declaração de Governo Aberto, e a aprovação da Lei de nº 12.527, de 18 de novembro de 2011, Lei de Acesso à Informação, marco regulatório fundamental para a consolidação da transparência pública no país (Brasil, 2011).

Apesar dos avanços normativos e institucionais, o processo de abertura de dados governamentais no Brasil ainda enfrenta desafios significativos. No âmbito dos Dados Governamentais Abertos (DGA), estudos internacionais e nacionais evidenciam fatores que podem impactar o processo, tais como a heterogeneidade dos usuários, os problemas técnicos no tratamento das informações, a qualidade e o formato dos dados disponibilizados (Davies; Bawa, 2012; Dawes; Helbig, 2010; Jassen; Charalabidis; Zuiderwijk, 2012; Zuiderwijk *et al.*, 2012; Sant'Ana, 2016).

Tais obstáculos contribuem de forma direta para a instauração de um estado de anomia, definido por Aleixo (2020, p. 205) como “[...] a ausência do reconhecimento e da prática das normas que se voltam à elaboração de estruturas que regulamentem um

³ A OGP é uma organização formada por líderes de diversos países e membros da sociedade civil que visa promover o acesso às informações públicas, ou seja, volta-se ao Governo Aberto. Conforme definido por Albano (2014, p. 23) “governo aberto é [a] disponibilização de informações em qualquer formato por parte dos governos e outras ações que visem promover maior transparência.”

processo de ações, que visam uma padronização, para incluir e estimular a participação da sociedade.”

Com o intuito de mitigar essas fragilidades, o governo brasileiro publicou, em 2012, o primeiro Plano de Ação da Infraestrutura Nacional de Dados Abertos (INDA)⁴. Elaborado pela Diretoria de Sistemas de Informação de Secretaria de Logística e Tecnologia da Informação tinha por propósito atender três finalidades: garantir a transparência na publicação dos DGA, orientar de forma estratégica e política a abertura dos dados na administração pública de 2013 a 2014, e subsidiar a criação de Planos de Publicação de Dados Abertos na INDA (Brasil, 2012a).

Contudo, como destaca Aleixo (2020) a efetividade de tais planos depende intrinsecamente do alinhamento com os “8 Princípios de dados abertos” – Princípios de Sebastopol. A ausência de alinhamento pode comprometer a padronização necessária e enfraquecer o papel do plano enquanto instrumento orientador das políticas de dados abertos.

Nesse contexto, este estudo tem como objetivo atualizar o diagnóstico realizado por Aleixo (2020) sobre a implementação dos princípios de dados abertos nos Planos de Ação da INDA referentes aos períodos de 2013-2014 e 2017-2018. Para tanto, incorpora-se também a análise das versões dos planos publicadas para os períodos de 2021-2022 e 2023-2025, a fim de identificar eventuais mudanças e atualizações, contribuindo para uma compreensão aprofundada dos avanços e desafios enfrentados na consolidação dos dados abertos no Brasil.

Esta pesquisa busca contribuir para a compreensão crítica dos desafios e avanços na implementação dos dados abertos no Brasil, oferecendo subsídios teórico-práticos ao aprimoramento das políticas públicas, com vistas ao fortalecimento da transparência e da participação cidadã.

⁴ A INDA é o conjunto de padrões, tecnologias e orientações para disseminação e compartilhamento de dados e informações públicas em formato aberto. Disponível em: <https://dados.gov.br/dados/conteudo/politica-de-dados-abertos#:~:text=O%20que%20%C3%A9%20a%20INDA,informa%C3%A7%C3%B5es%20p%C3%BAblicas%20em%20formato%20aberto>. Acesso em: 01 fev. 2025.

2 PROCEDIMENTOS METODOLÓGICOS

Trata-se de uma pesquisa qualitativa, com caráter exploratório, descritivo e aplicado. A coleta dos documentos foi realizada na Base de Conhecimento da Controladoria Geral da União (CGU)⁵. A pesquisa analisou os seguintes Planos de Ação:

- a) Plano de Ação da INDA 2013-2014 (Brasil, 2012a);
- b) Plano de Ação da INDA 2017-2018 (Brasil, 2017);
- c) Plano de Ação da INDA 2021-2022 (Brasil, 2021);
- d) Plano de Ação da INDA 2023-2025 (Brasil, 2023).

Os dados extraídos dos Planos de Ação da INDA foram sistematizados por meio de análise documental categorizada, com base nos “8 Princípios de dados abertos” – Princípios de Sebastopol - citados como referência na formulação da INDA no Brasil. A abordagem permitiu identificar a evolução regulatória e atender ao objetivo da pesquisa.

3 RESULTADOS

A INDA visa padronizar a gestão e a disseminação de dados públicos, por meio de um catálogo governamental, incentivando sua utilização e agregação de valor pela sociedade (Aleixo, 2020). A Figura 1 apresenta sua linha do tempo com os principais eventos e ações.



Fonte: Elaboração própria.

⁵ Disponível em: <https://repositorio.cgu.gov.br/handle/1/69418>. Acesso em: 01 fev. 2025.

Desde 2012, a INDA tem se aprimorado por meio do fortalecimento da governança, ampliação da participação social, incorporação de tecnologias e alinhamento a iniciativas globais. O Quadro 1 sintetiza suas principais ações e impactos na sociedade.

Quadro 1 - Evolução dos Planos de Ação da INDA: foco, ações, diretrizes e impactos

PLANO DE AÇÃO	ANO DE PUBLICAÇÃO	FOCO	AÇÕES	DIRETRIZES	IMPACTOS ESPERADOS
2013-2014	2012	Transparência governamental, padronização e estruturação dos dados	Publicação dos dados	Acesso e conformidade com padrões de dados	Aumento da transparência e acesso à informação pública
2017 - 2018	2017	Refinamento de dados abertos e governança	Aperfeiçoamento da estrutura e das ferramentas	Melhoria da qualidade e transparência dos dados	Melhoria na qualidade dos dados e maior conformidade
2021 - 2022	2021	Revisão e melhorias da infraestrutura de dados	Atualização de dados e ampliação da transparência	Foco em qualidade e governança de dados públicos	Avanço no uso de dados abertos e melhor governança
2023 - 2025	2023	Fortalecimento das políticas de dados abertos	Expansão da governança e maior acesso aos dados	Melhoria contínua e conformidade internacional	Maior colaboração internacional e adaptação às normas globais

Fonte: Elaboração própria a partir do Repositório da CGU (Brasil, 2025).

O quadro evidencia um esforço contínuo de aprimoramento dos Planos de Ação da INDA, orientado ao fortalecimento das políticas públicas, ao aprimoramento da governança dos dados e a conformidade com normas internacionais. Observa-se, contudo, uma descontinuidade no período de 2014 a 2016, sem registros oficiais que justifiquem a interrupção. Presume-se que fatores políticos e institucionais tenham contribuído para essa lacuna, sendo relevante destacar que o Decreto de nº 8.777, de 11 de maio de 2016, ao instituir a Política de Dados Abertos do Poder Executivo Federal (Brasil, 2016), sinaliza a retomada formal das iniciativas voltadas aos dados abertos.

As melhorias observadas em aspectos como qualidade, conformidade e acesso aos dados ao longo dos anos, conforme descrito nos documentos, sinaliza o esforço do Estado para ampliar a transparência e consolidar as infraestruturas voltadas à abertura e disseminação dos dados governamentais.

Os documentos orientadores que ampararam a implantação da INDA referem-se à sua estrutura como baseada nos Princípios de transparência do Reino Unido, os quais estão direcionados ao conceito de governo aberto e envolvem três principais atores: o Estado, as tecnologias e os usuários (Harrison, Pardo e Cook, 2012). No entanto, estes documentos não fazem menção direta aos princípios de Dados abertos, como os "8 Princípios de Dados Abertos"⁶.

As versões mais recentes do Plano de Ação da INDA, adotam uma nova estrutura, substituindo diretrizes e objetivos gerais por ações concretas, com entregas, prazos e responsáveis definidos (Brasil, 2025b). Tal reformulação pode indicar uma mudança na abordagem da política de dados abertos no Brasil, de um modelo normativo e conceitual para uma estrutura mais operacional, focado na implementação e no monitoramento, como evidenciado no detalhamento das ações previstas no Plano de Ação 2023–2025⁷.

No Plano de Ação (2021-2022) são relacionadas 10 (dez) ações, enquanto na versão 2023-2025 são apresentadas 7 (sete) ações (Quadro 2).

Quadro 2 – Comparativo das ações dos Planos de Ação da INDA (2021-2022 e 2023-2025)

AÇÕES	PLANO DE AÇÃO - 2021-2022	PLANO DE AÇÃO - 2023-2025
1	Promover a abertura de bases de alto valor	Workshop de dados governamentais abertos
2	Aprimorar o Framework de Dados Abertos para o governo federal	Promoção da inovação com dados abertos
3	Revisão geral das bases de dados publicadas no Portal Brasileiro de Dados Abertos	Programa continuado de orientação e capacitação em dados abertos
4	Realizar pesquisas sobre abertura e reuso de dados	Realizar pesquisa sobre reutilização de dados governamentais abertos
5	Realizar eventos de cocriação com o uso de dados abertos	Revisar framework orientativo sobre dados abertos
6	Programa continuado de orientação e capacitação em dados abertos e reuso de dados	Definir modelo de catalogação de conjuntos de dados (inventário de dados)
7	Reestruturação do Portal Brasileiro de Dados Abertos	Sistematização de recebimento e implementação de sugestões estruturais para correção, melhorias e validação de dados das bases de dados federais

⁶ Os “Princípios de Sebastopol” foram “elaborados em dezembro de 2007 por um grupo de defensores do movimento *Open Government* na Califórnia, sob a coordenação de Tim O’Reilly e Carl Malamud, os quais viriam a orientar os dados abertos e a elaboração de políticas com vistas às iniciativas dos dados governamentais abertos em diversos países” (Aleixo, 2020, p. 123).

⁷ Disponível em: disponível em: www.dados.gov.br.

8	Catálogo de Bases de Dados Federais	-----
9	Fomento a adesões de entes federativos e de órgãos dos demais Poderes da República ao Portal Brasileiro de Dados Abertos	-----
10	Normativo de reestruturação da Infraestrutura Nacional de Dados Abertos - INDA	-----

Fonte: Baseado em Aleixo (2020, p. 192-193).

Ao analisar as ações indicadas nos Planos, observa-se uma reorganização que evidencia uma nova ênfase em aspectos como inovação, capacitação e aprimoramento técnico. No entanto, nota-se que algumas ações deixaram de constar na versão mais recente, como o fomento à adesão de entes federativos e a reestruturação da INDA. A ausência pode ser vista com preocupação, pois pode representar um recuo na expansão e cobertura e do acesso aos dados governamentais abertos, bem como uma menor ênfase na coordenação, padronização e integração de esforços em âmbito nacional.

Aleixo (2020), ao analisar as versões iniciais dos Planos de Ação da INDA à luz dos “8 Princípios de dados abertos” evidenciou avanços na estruturação e governança dos dados, mas também apontou lacunas, como a ausência de diretrizes claras para os princípios de não propriedade, não discriminação e licenciamento aberto. Ao atualizar essa análise com os Planos de Ação da INDA dos períodos de 2021-2022 e 2023-2025, observa-se a seguinte configuração (Quadro 3).

A análise da evolução dos Planos de Ação da INDA, revela um movimento gradual, porém ainda incipiente, de incorporação dos princípios que fundamentam a política de dados abertos no Brasil.

Conforme assinala Aleixo (2020), o primeiro ciclo (2013-2014) caracteriza-se pela ausência de diretrizes consolidadas, refletindo um estágio inicial da política. No ciclo (2017-2018), embora sejam identificados avanços normativos com a incorporação de princípios anteriormente não contemplados; a retirada de princípios já presentes revela discontinuidades e fragilidades conceituais no desenvolvimento da política.

Nos planos subsequentes, observa-se um processo de amadurecimento institucional, com a incorporação de ações voltadas à capacitação, à inovação e ao fortalecimento da governança. Tal movimento indica uma evolução positiva no

tratamento da temática, ainda que apresente limitações que comprometem a consolidação dos princípios de dados abertos como eixo estruturante da política.

Quadro 3 – Comparativo entre os Planos de Ação da INDA e a observância aos Princípios

PRINCÍPIO ⁸	PLANO DE AÇÃO (2013-2014)	PLANO DE AÇÃO (2017-2018)	PLANO DE AÇÃO 2021-2022	PLANO DE AÇÃO 2023-2025
Completo⁹	Não contempla	I. observância da publicidade das bases de dados como preceito geral e do sigilo como exceção;	Ações 1, 2, 3, 7, e 8	Ações 5, 6 e 7
Primário¹⁰	Não contempla	V. completude e interoperabilidade das bases de dados, as quais devem ser disponibilizadas em sua forma primária, com o maior grau de granularidade possível, ou referenciar as bases primárias, quando disponibilizadas de forma agregada;	Ações 1 e 2	Não contempla
Atual¹¹	V. Os dados públicos devem ser publicados com rapidez e precisão VI. Publicar dados rapidamente, e então republicá-lo como <i>linked data</i>	VI. atualização periódica, de forma a garantir a perenidade dos dados, a padronização de estruturas de informação e o valor dos dados à sociedade e atender às necessidades de seus usuários; e	Ações 1, 2, 3 e 7	Ação 1

⁸ Para auxiliar na compreensão de cada princípio as definições foram inseridas como nota de rodapé.

⁹ Todos os dados públicos são disponibilizados, sem restrições indevidas de privacidade, segurança ou controle de acesso (Aleixo, 2020).

¹⁰ Dados coletados na origem, com o nível mais alto possível de granularidade, sem agregações ou modificações (Aleixo, 2020).

¹¹ Disponibilizados o mais rápido possível para preservar seu valor (Aleixo, 2020).

Acessível¹²	IV. Os dados públicos estarão disponíveis e fáceis de encontrar através de um único e fácil ponto de acesso on-line (dados.gov.br)	IV. permissão irrestrita de reuso das bases de dados publicadas em formato aberto;	Ações 1, 2, 3, 4, 5, 6, 7, 8 e 9	Todas as ações
Legível por máquinas¹³	II. Os dados públicos serão publicados de forma reutilizável e legível por máquina	II. garantia de acesso irrestrito às bases de dados, as quais devem ser legíveis por máquina e estar disponíveis em formato aberto;	Ações 2 e 3	Ações 1, 2, 3 e 6
Não discriminatório¹⁴	VII. Os dados públicos estarão disponíveis gratuitamente para uso em qualquer forma legal	Não contempla	Ações 5, 6, 7, 8, 9 e 10	Ação 1
Não proprietários¹⁵	Não contempla	Não contempla	Ação 2	Ações 2, 3 e 5
Livre de licença¹⁶	III. Os dados públicos devem ser publicados sob licença aberta que permita a livre reutilização, incluindo a reutilização comercial	Não contempla	Ação 2	Ações 1, 2, e 3

Fonte: Baseado em Aleixo (2020, p. 192-193).

A partir de uma análise interpretativa das ações previstas no Plano 2021-2022, ainda que os princípios de dados abertos não sejam explicitamente mencionados, é possível identificar sua presença em diversas ações. Alguns princípios são atendidos de forma indireta, a partir da análise crítica das propostas. Ainda assim, a ausência de referências claras e diretas aos princípios evidencia uma fragilidade na consolidação da Política Nacional de Dados Abertos, especialmente no que se refere à sua institucionalização normativa.

¹² Abertos ao público mais amplo possível e para os mais variados propósitos (Aleixo, 2020).

¹³ Estruturados de forma a permitir o processamento automatizado (Aleixo, 2020).

¹⁴ Disponíveis para todos, sem necessidade de identificação ou registro (Aleixo, 2020).

¹⁵ Disponíveis em formatos abertos, sem controle exclusivo por qualquer entidade (Aleixo, 2020).

¹⁶ Não sujeitos a direitos autorais, marcas ou patentes, exceto em casos regulados por estatutos (Aleixo, 2020).

No Plano 2023-2025, apresenta um movimento ambíguo. Por um lado, amplia-se a cobertura dos princípios nas ações planejadas — com destaque para o princípio da acessibilidade, contemplado em praticamente todas as ações. Por outro, persistem lacunas importantes, como a ausência de normativas robustas que garantam a implementação efetiva de princípios como o da primariedade, da não propriedade e da liberdade de licença. Essa assimetria revela uma política que, embora mais madura institucionalmente, ainda enfrenta obstáculos estruturais e normativos.

Nesse sentido, compreende-se que o ciclo de amadurecimento institucional avança da formulação de diretrizes para a proposição de ações. No entanto, tais avanços não superam os desafios relacionados à articulação estratégica, à coerência entre os planos e à efetiva normatização dos princípios. Este olhar reforça a crítica de Aleixo (2020) quanto à fragmentação das ações e à ausência de uma trajetória consistente de desenvolvimento da política pública de dados abertos no país.

Assim, a trajetória dos Planos de Ação revela uma adesão progressiva, porém ainda insuficiente, aos princípios dos dados abertos. A incorporação fragmentada, a ausência de marcos regulatórios claros e a falta de coerência interna entre os planos sugerem a necessidade urgente de revisão e fortalecimento da política pública de dados abertos no Brasil, de modo a garantir sua plena efetividade e alinhamento com os padrões internacionais de governo aberto.

4 CONSIDERAÇÕES FINAIS

Conforme percorrido ao longo do trabalho, a implementação de políticas de dados abertos tem sido um desafio contínuo em vários países, e sobretudo no Brasil. A INDA, por meio de seus Planos de Ação, tem assumido o papel de direcionar as ações voltadas à essa agenda.

O objetivo deste trabalho foi atualizar o diagnóstico realizado anteriormente sobre a implementação dos princípios de dados abertos nos Planos de Ação da INDA de 2013-2014 e 2017-2018, incorporando a análise das versões de 2021-2022 e 2023-2025, para identificar possíveis mudanças e as atualizações.

Os resultados indicam que, embora haja avanços pontuais na adoção de algumas práticas, como a acessibilidade e a atualidade dos dados, ainda persistem lacunas significativas na aplicação plena dos princípios dos dados abertos. Aspectos como a participação efetiva da sociedade civil e o incentivo ao reuso dos dados continuam sendo áreas críticas. A evolução nos planos mais recentes revela um esforço em melhorar a qualidade e o alcance dos dados, mas a infraestrutura necessária para garantir a efetiva implementação de todos os princípios ainda está em desenvolvimento.

Conclui-se, portanto, que a política de dados abertos no Brasil, ainda existem desafios significativos que comprometem a plena implementação dos princípios de dados abertos. A crescente dependência de dados de qualidade destaca a importância de garantir dados acessíveis e confiáveis, fundamentais para o desenvolvimento de tecnologias emergentes, como a inteligência artificial (IA).

No entanto, a persistente ausência ou o distanciamento dos princípios nos documentos reguladores e orientadores do Estado resultam em um cenário de anomia no acesso aos dados governamentais abertos, enfraquecendo seu potencial como instrumento de controle social e inovação. Este contexto, reforça a urgência de ações concretas para assegurar que os dados abertos no Brasil sejam, de fato, acessíveis, utilizáveis e alinhados às diretrizes internacionais para a participação cidadã.

REFERÊNCIAS

- ALBANO, C. S. **Dados governamentais abertos**: proposta de um modelo de produção e utilização sob a ótica conceitual da cadeia de valor. 2014. Tese (Doutorado em Administração) - Faculdade de Economia, Administração e Contabilidade, Universidade de São Paulo, São Paulo, 2014. Disponível em: <http://www.teses.usp.br/teses/disponiveis/12/12139/tde-03062014-170642/publico/ClaudioSonaglioAlbanoVC.pdf>. Acesso em: 19 jun. 2019.
- ALEIXO, D. V. B. **O estado de anomia dos dados no acesso aos dados governamentais abertos no Brasil**. 2020. Tese (Doutorado em Ciência da Informação) - Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2020. Disponível em: <https://repositorio.unesp.br/server/api/core/bitstreams/21216708-7e2b-4838-ab11-8d5683b1cba0/content>. Acesso em: 26 fev. 2025.

BRASIL. **Conteúdo:** plano de ação da INDA 2021-2022. 2025b. Disponível em: <https://dados.gov.br/dados/conteudo/plano-de-acao-da-inda-2021-2022>. Acesso em: 26 fev. 2025.

BRASIL. Controladoria Geral da União. **Repositório de Conhecimento da CGU.** 2025a. Disponível em: <https://repositorio.cgu.gov.br/handle/1/69418>. Acesso em: 01 fev. 2025.

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:** seção 1, Brasília, DF, 12 maio 2016. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2016/decreto/d8777.htm. Acesso em: 01 fev. 2026.

BRASIL. Instrução Normativa nº 4. Institui a Infraestrutura Nacional de Dados Abertos – INDA. 2012a. **Diário Oficial da União:** seção 1, Brasília, DF, 13 abr. 2012. Disponível em: <https://www.gov.br/governodigital/pt-br/dados-abertos/InstrucaoNormativaINDA42012.pdf>. Acesso em: 01 fev. 2025.

BRASIL. Lei nº 12.527 de 18 de novembro de 2011. Regula o acesso a informações previsto no inciso XXXIII do art. 5º, no inciso II do § 3º do art. 37 e no § 2º do art. 216 da Constituição Federal; altera a Lei nº 8.112, de 11 de dezembro de 1990; revoga a Lei nº 11.111, de 5 de maio de 2005, e dispositivos da Lei nº 8.159, de 8 de janeiro de 1991; e dá outras providências. **Diário Oficial da União:** seção 1, Brasília, DF, 18 nov. 2011. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm. Acesso em: 01 fev. 2026.

BRASIL. Ministério do Planejamento, Desenvolvimento e Gestão. Secretaria de Tecnologia da Informação. **Plano de ação da Infraestrutura Nacional de Dados Abertos 2017-2018.** 2017. Disponível em: <https://repositorio.cgu.gov.br/handle/1/77207>. Acesso em: 01 fev. 2025.

BRASIL. **Plano de ação da Infraestrutura Nacional de Dados Abertos 2023-2025.** 2023. Disponível em: <https://repositorio.cgu.gov.br/handle/1/77207>. Acesso em: 01 fev. 2025.

BRASIL. **Plano de ação da Infraestrutura Nacional de Dados Abertos 2021-2022.** 2021. Disponível em: <https://repositorio.cgu.gov.br/handle/1/77207>. Acesso em: 01 fev. 2025.

BRASIL. Sistema de Administração dos Recursos de Informação e Informática. **Estratégia Geral de Tecnologia da Informação (2011-2012).** 2011. Disponível em: https://www.gov.br/governodigital/pt-br/estrategia-de-governanca-digital/Estrategia_Geral_de_TI_2011_2012_SISP.pdf. Acesso em: 01 fev. 2025.

DAVIES, T. G.; BAWA, Z. A. The promises and perils of.opengovernment data (OGD). **The Journal of Community Informatics**, v. 8, n. 2, p. 1-6, 2012.

DAWES, S. S.; HELBIG, N. Information strategies for open government: challenges and prospects for deriving public value from government transparency. *In*: INTERNATIONAL CONFERENCE ON ELECTRONIC GOVERNMENT, 2010, Berlin. **Anais [...]**. Berlin: Springer,

2010. p. 50-60. Disponível em: https://link.springer.com/chapter/10.1007/978-3-642-14799-9_5. Acesso em: 23 maio 2025.

HARRISON, T.; PARDO, T.; COOK, M. Creating open government ecosystems: a research and development agenda. **Future Internet**, v. 4, p. 900-928, 2012.

JANSSEN, M.; CHARALABIDIS, Y.; ZUIDERWIJK, A. Benefits, adoption barriers and myths of open data and open government. **Information Systems Management**, v. 29, n. 4, p. 258-268, 2012.

SANT'ANA, R. C. G. Ciclo de vida dos dados: uma perspectiva a partir da Ciência da Informação. **Informação & Informação**, Londrina, v. 21, n. 2, p. 116-142, dez. 2016. Disponível em: <http://www.uel.br/revistas/uel/index.php/informacao/article/view/27940/20124>. Acesso em: 26 fev. 2025.

ZUIDERWIJK, A. *et al.* Socio-technical impediments of Open Data. **Electronic Journal of e-Government**, v. 10, n. 2, 2012.



[4.0 Internacional](https://creativecommons.org/licenses/by/4.0/)

Esta licença permite compartilhamento, remixe, adaptação e criação a partir do trabalho, mesmo para fins comerciais, desde que sejam atribuídos créditos ao(s) autor(es). Conteúdos elaborados por terceiros, citados e referenciados nesta obra não são cobertos pela licença.

Como citar este trabalho:

ALEIXO, Diana Vilas Boas Souto; SANT'ANA, Ricardo Cesar Gonçalves. Desafios e avanços na implementação dos princípios de dados abertos no Brasil: uma análise dos planos de ação da INDA. *In*: WORKSHOP DE INFORMAÇÃO DADOS E TECNOLOGIA, 8., 2025, Marília, SP. **Anais [...]**. Marília, SP: Universidade de Marília, 2025. DOI: <https://doi.org/10.22477/viii.widat.230>.



WIDaT 2025

VIII WORKSHOP DE INFORMAÇÃO,
DADOS E TECNOLOGIA

MARÍLIA • SP

Catálogo Nacional de Dados do Governo Federal: análise da utilização do vocabulário DCAT-BR

National Data Catalogue of the Federal Government: analysis of the use of DCAT-BR vocabulary

Katyusha Madureira Loures de Souza¹



<http://lattes.cnpq.br/9649842485745030>



<https://orcid.org/0000-0002-0617-8518>

Mariana Baptista Brandt²



<http://lattes.cnpq.br/3761037263199030>



<https://orcid.org/0000-0001-8119-7527>

Silvana Aparecida Borsetti Gregorio Vidotti³



<http://lattes.cnpq.br/7390573927636069>



<https://orcid.org/0000-0002-4216-0374>

Resumo

O Catálogo Nacional de Dados (CND) foi criado pelo governo brasileiro com o objetivo de facilitar a descoberta de dados e promover a interoperabilidade de dados abertos entre órgãos públicos. Baseado no *Data Catalog Vocabulary* (DCAT) da W3C, o CND estrutura a descrição de conjuntos de dados na web com a especialização DCAT-BR, criada para este fim. Este artigo analisa a implementação do DCAT-BR no CND e sua eficácia para alcançar os objetivos de descoberta de dados e interoperabilidade. A pesquisa documental e bibliográfica revelou que a descoberta de dados é apenas parcialmente atendida, devido à falta de descrições detalhadas das variáveis dos conjuntos de dados. Em relação à interoperabilidade, identificou-se que o DCAT-BR não cobre todos os metadados necessários e apresenta redundâncias na relação entre alguns termos e metadados. Conclui-se, portanto, que são necessárias melhorias para que o CND cumpra seus objetivos de maneira mais abrangente.

Palavras-chave: dados abertos governamentais; catálogo de dados; vocabulário semântico; interoperabilidade; descoberta de dados.

Abstract

The National Data Catalog (CND) was created by the Brazilian government to facilitate data discovery and promote the interoperability of open data among public agencies. Based on the Data Catalog Vocabulary (DCAT)

¹ Mestre em Ciência da Informação, Universidade de Brasília (UnB), Brasília, DF, Brasil. Analista em Ciência e Tecnologia, Fundação Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), DF, Brasil. katyusha.souza@unesp.br.

² Doutora em Ciência da Informação, Universidade Estadual Paulista (Unesp), Marília, SP, Brasil. Analista legislativa, Câmara dos Deputados, Brasília, DF, Brasil. mariana.brandt@unesp.br.

³ Doutora em Educação, Universidade Estadual Paulista Júlio de Mesquita Filho (Unesp), Marília, São Paulo, Brasil. Professora do Departamento e do Programa de Pós-graduação em Ciência da Informação, Universidade Estadual Paulista Júlio de Mesquita Filho (Unesp), Marília, São Paulo, Brasil. Coordenadora de Tecnologias Aplicadas (COTEA), Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), Brasília, Distrito Federal, Brasil. silvana.vidotti@unesp.br.

from W3C, the CND structures the description of datasets on the web using DCAT-BR, a specialization created for this purpose. This article analyzes the implementation of DCAT-BR in the CND and its effectiveness in achieving the goals of data discovery and interoperability. Documentary and bibliographic research revealed that data discovery is only partially met due to the lack of detailed descriptions of dataset variables. Regarding interoperability, it was identified that DCAT-BR does not cover all necessary metadata and presents redundancies in the relationship between some terms and metadata. Therefore, it is concluded that improvements are needed for the CND to fully achieve its objectives in a more comprehensive manner.

Keywords: open government data; data catalog; semantic vocabulary; interoperability; data discovery.

1 INTRODUÇÃO

A disponibilização de dados abertos governamentais tem crescido nos últimos anos, especialmente após a promulgação da Lei de Acesso à Informação (LAI), que determina que os órgãos públicos devem divulgar as informações em seus sites de modo a “[...] possibilitar o acesso automatizado por sistemas externos em formatos abertos, estruturados e legíveis por máquina (Brasil, 2011)”. O governo brasileiro conta com um portal de dados abertos (dados.gov.br) que armazena 14.209⁴ conjuntos de dados de 269 órgãos da administração pública. Além disso, vários órgãos publicam conjuntos de dados abertos de sua responsabilidade na própria página web do órgão.

Em dezembro de 2024 foi lançada uma nova ferramenta para disponibilização dos dados, o Catálogo Nacional de Dados (CND), com objetivos de descoberta de dados e interoperabilidade entre os órgãos (Brasil, 2024b). O CND utiliza o vocabulário semântico *Data Catalog Vocabulary* (DCAT), do *World Wide Web Consortium* (W3C), para descrição de conjuntos de dados, com uma especialização para o governo brasileiro denominada DCAT-BR.

A descoberta de dados é um objetivo alinhado aos objetivos bibliográficos, ou seja, os objetivos de um catálogo, que são: encontrar, identificar, selecionar, obter e navegar (IFLA, 2018). O objetivo deste artigo é analisar a implementação do DCAT no CND e sua potencialidade para atingir os objetivos de descoberta e de interoperabilidade dos dados propostos pela ferramenta. Para isso, foi realizada uma pesquisa documental nas fontes governamentais disponíveis sobre a elaboração do CND, nos portais de dados abertos dos órgãos e na documentação técnica do DCAT, além de pesquisa bibliográfica para referencial teórico.

⁴ Consulta em 18 de janeiro de 2025.

2 O CATÁLOGO E SUAS FUNÇÕES

Por definição, um catálogo deve ser um instrumento eficaz e efetivo que permita ao usuário **encontrar** recursos bibliográficos numa coleção como resultado de uma pesquisa, utilizando atributos e relações entre recursos; **identificar** um recurso bibliográfico ou agente; **selecionar** um recurso bibliográfico que seja apropriado às necessidades do usuário; adquirir ou **obter** acesso a um item descrito e **navegar e explorar** dentro de um catálogo ou para além dele (IFLA, 2018). A *International Federation of Library Associations and Institutions* (IFLA) utiliza sempre o termo “recurso bibliográfico” devido a estar inserida no contexto bibliográfico, por ser uma Associação de Bibliotecas. Porém, os recursos descritos em um catálogo podem ser de outras naturezas sem alteração de suas funções e objetivos.

Em seu trabalho de 2008, Cunha e Cavalcanti, reuniram diversas definições sobre o termo ‘catálogo’ que abordam desde conceituações envolvendo a tipologia do documento - documento (secundário, lista, relação), função (analítico, entrada catalográfica), organização (alfabético, classificado por assunto, por autor, sistemático, analítico, assindético, bibliográfico, automatizado), por acesso (aberto, central, coletivo) e por função (comercial, de aquisição, de autoridade, dentre outros). À época, os autores já mencionavam o catálogo de dados com remissiva para o dicionário de dados (Cunha; Cavalcanti, 2008, p.73). Para esse termo, os autores propuseram a seguinte definição:

Inf 1. Programa associado com o sistema de gerenciamento de banco de dados e que mostra as definições dos dados usados pelo sistema. Funciona como auxiliar para a pessoa responsável pela administração e manutenção do banco de dados. 2. Listagem organizada de todos os elementos pertinentes ao sistema (Cunha; Cavalcanti, 2008, p. 73).

Os termos “catálogo de dados” e “dicionários de dados”, tratados como sinônimos à época, assumem características diferentes a partir dos avanços na produção, armazenamento e acesso a dados (big data). Diferente do dicionário de dados, que descreve dados de tabelas de bancos de dados internos das organizações e possuem função de documentação de sistemas, os catálogos de dados são ferramentas abrangentes para descrição de conjuntos de dados diversos, que podem ser de diferentes instituições e que são tratados aqui como recursos informacionais. Os catálogos de dados devem, assim, assumir as mesmas funções de um catálogo bibliográfico, permitindo aos usuários encontrar, identificar, selecionar e obter

os conjuntos de dados que descrevem, além de permitir a navegação, explorando as possibilidades de descoberta de dados no catálogo.

3 DADOS ABERTOS GOVERNAMENTAIS

Os dados produzidos pelos órgãos do governo são de interesse público e podem ser usados para vários fins. A transparência das informações é necessária para a manutenção dos governos democráticos e a publicação dos dados governamentais em formato aberto é uma das formas de viabilizar essa transparência. Assim, os vários órgãos publicam seus dados em seus próprios repositórios ou em repositórios coletivos. Esses repositórios necessitam de um cadastro dos dados publicados, ou seja, de um catálogo que traga informações sobre os dados, para que esses possam ser representados e encontrados.

Assim, o catálogo de dados torna-se um instrumento necessário para viabilizar a transparência ativa por meio da publicação dos dados produzidos: “Catálogos de Dados são ferramentas de gestão de metadados que apoiam a curadoria de dados oferecendo a capacidade de inventariar e descobrir dados em uma plataforma integrada, conectando a oferta com a demanda de dados”. (Jahnke; Otto, 2023, p. 90, tradução nossa).

Ou seja, para os autores o propósito fundamental é o inventariado e a descoberta desses dados. Além disso, os autores também mencionam outras possibilidades identificadas durante sua pesquisa, como a avaliação de dados (uso, qualidade, perfil), governança de dados (fluxos, papéis e responsabilidades, regras e políticas), colaboração de dados (*tagging*, compartilhamento, comentários) e análise de metadados (estória de dados, repositórios de aplicação de dados). Ademais, suas capacidades funcionais facilitam o apoio à visualização, administração e automação, incluindo características da inteligência artificial.

Como relatado na seção anterior, os catálogos buscam cumprir objetivos relacionados ao seu uso, em especial o de encontrar itens (livros, documentos, dados) necessários para o usuário. Esse é também o objetivo do Catálogo Nacional de Dados, quando se refere ao objetivo de descoberta de dados. O CND, objeto deste estudo, será descrito a seguir.

3.1 Catálogo Nacional de Dados

O Catálogo Nacional de Dados é um projeto realizado a partir de uma parceria entre o Ministério da Gestão e da Inovação em Serviços Públicos (MGI) e a Controladoria-Geral da

União (CGU). Segundo o Guia rápido do Catálogo Nacional de Dados (Brasil, 2024c), o CND “é um instrumento para apoiar os órgãos na organização e disponibilização das informações sobre os dados em posse do governo a fim de que se tornem úteis para todos” e tem por objetivo “facilitar a descoberta e promover o reuso dos dados produzidos pelo governo”. O instrumento foi criado para identificar com clareza os dados existentes no governo, tornar os dados mais acessíveis, incentivar o uso e reuso de dados, transformar os dados em ativos informacionais e desenvolver produtos e políticas públicas (Brasil, 2024b).

Neste modelo, são catalogados os conjuntos de dados e os recursos em 3 níveis de catalogação, a depender dos metadados utilizados. Esses níveis são denominados “selos”, nas categorias ouro, prata e bronze. “Os conjuntos de dados recebem selos de acordo com a quantidade de campos preenchidos e a validação negocial dos campos. Total de 44 campos.” (Brasil, 2024a). A ideia é implementar os selos de acordo com um cronograma, que começa com o selo bronze e prevê a implementação do selo prata até outubro de 2025 e do selo ouro até abril de 2026. São esperados diversos benefícios advindos do catálogo, como o uso de dados para desenvolvimento de novos produtos e serviços, políticas públicas e para pesquisa e desenvolvimento. Segundo a cartilha, o catálogo “é uma ferramenta poderosa que pode ajudar o governo federal a melhorar seus serviços e a tomar decisões mais informadas.” (Brasil, 2024a, p. 6).

No dia 20 de dezembro de 2024, a Controladoria Geral da União (CGU), segundo matéria publicada em sua página eletrônica⁵, disponibilizou uma nova versão do Portal Brasileiro de Dados abertos com modificações para aumentar o grau de transparência e participação social a partir da incorporação das funcionalidades do Catálogo Nacional de Dados. Segundo a notícia publicada, o objetivo seria facilitar a descoberta de dados e promover a interoperabilidade entre os órgãos públicos que alimentam o portal.

Essa facilidade seria proporcionada pela catalogação também de conjuntos de dados não abertos, que segundo a CGU “promoverá maior transparência pública aos dados de posse dos órgãos públicos federais” e “trará mais robustez ao processo de elaboração dos Planos de Dados Abertos, pois contribuirá com a sua continuidade, ao registrar digitalmente o inventário de dados da Administração Pública”.

⁵ Disponível em: <https://www.gov.br/cgu/pt-br/assuntos/noticias/2024/12/cgu-lanca-novas-funcionalidades-no-portal-de-dados-abertos>.

A notícia destaca melhorias no portal de dados, incluindo a criação do selo Bronze, aplicado a conjuntos de dados que atendem aos requisitos mínimos do padrão DCAT-BR Bronze, por meio de um *endpoint* de *harvesting* que busca RDFs diretamente das URLs de órgãos públicos que não utilizam CKAN. Outras melhorias incluem novos campos no detalhamento dos conjuntos de dados, checagem periódica de atualizações, identificação de dados sensíveis sobre raça/etnia para apoiar políticas públicas de ações afirmativas e aprimoramento da catalogação, com sinalização sobre a abertura dos dados, questões legais, relação com os Objetivos de Desenvolvimento Sustentável (ODS) e direcionamento para dados relacionados.

A atualização também incluiu a possibilidade de indicação da área técnica responsável pela catalogação (com respectivo e-mail, bem como sua versão e a existência do atributo de gênero). O cadastro da organização para publicação no Portal de Dados Abertos inclui o contato com a CGU e o preenchimento de um formulário com 18 campos (13 obrigatórios), que inclui o cadastro no Sistema de Organização e Inovação Institucional do Governo Federal (SIORG)⁶. Após o cadastro da organização também é necessário cadastrar o gestor da organização que ficará responsável pelo cadastramento dos dados. A indicação também deverá ser realizada através do contato com a CGU.

O cadastro de dados da organização pode ser realizado em duas etapas: conjunto de dados e recursos (que estão vinculados ao conjunto cadastrado). Na primeira etapa, é preciso informar o título do conjunto de dados, observância legal, descrição, licença de uso, área técnica responsável, periodicidade de atualização, tema, relacionamento com os Objetivos de Desenvolvimento Sustentável (uma parceria do Brasil com a ONU, lembrando também a Agenda 2030), se o conjunto possui dados sensíveis, palavras-chave que descrevem o conjunto de dados, sua cobertura temporal, espacial, versão (inclusive se foi descontinuado), visibilidade e associações (interoperabilidade). Na segunda etapa, referente aos recursos, basta informar o título do recurso, URL de acesso, tipo do recurso, formato e uma descrição.

O dicionário do que significa cada campo para que o usuário possa esclarecer dúvidas quanto ao seu preenchimento está disponível em um material do tipo "saiba mais" na página do CDN, no entanto não se identificou *tooltips* na tela de cadastramento do material.

⁶ Saiba como publicar um conjunto de dados. Disponível em: <https://dados.gov.br/dados/conteudo/saiba-como-publicar-um-conjunto-de-dados>. Acesso em 24 fev. 2025.

3.2 Vocabulários Semânticos para Descrição de Dados: DCAT e DCAT-BR

Em sua página, a W3C (2024) informa que o *Data Catalog Vocabulary* (DCAT), que está em sua 3a. versão, é um vocabulário do RDF desenvolvido para facilitar a interoperabilidade entre catálogos de dados publicados na web, na qual define um modelo e fornece exemplos para seu uso. A interoperabilidade pode ser organizacional, técnica ou semântica, e se refere a capacidade de trabalho conjunto entre sistemas e organizações garantindo que pessoas, organizações e sistemas computacionais possam interagir para trocar informações de maneira eficaz e eficiente (Brasil, 2012). No caso da interoperabilidade entre catálogos, trata-se da interoperabilidade semântica, que se refere

[...] a capacidade de dois ou mais sistemas heterogêneos e distribuídos trabalharem em conjunto, compartilhando as informações entre eles com entendimento comum de seu significado (Buranarach, 2004). A interoperabilidade semântica garante que os dados trocados tenham seu efetivo significado corretamente interpretado dentro do contexto de uma dada transação ou busca de informação, dentro da cultura, convenções e terminologias adotadas por cada setor ou organização e, assim, compartilhados pelas partes envolvidas. (Brasil, 2012, p.8)

Assim, segundo o W3C, o DCAT permite a um editor descrever conjuntos de dados e serviços de dados em um catálogo usando modelos padronizados e vocabulário que facilite o consumo e unificação de metadados de diversos catálogos. A solução tem como proposta tornar possível a publicação de catálogos de dados de forma descentralizada. A versão 3 deste catálogo buscou acrescentar casos de uso, requisitos e a experiência da comunidade envolvida (*Data eXchange Working Group, s.i.c.*) apoiando abordagens relacionadas à descrição e intercâmbio de dados.

A solução foi originalmente desenvolvida e hospedada na *Digital Enterprise Research Institute* (DERI), um fundo mantido pela *Science Foundation Ireland* (SFI) que era responsável pelo plano nacional de desenvolvimento em estratégias para ciência tecnologia e inovação no período de 2007 a 2013. Em 2014, o *eGov Interest Group* passou a trabalhar no refino e depois padronizado pelo *Government Linked Data (GLD) Working Group*, ambos do W3C. Sua segunda versão foi desenvolvida pela *Dataset Exchange Working Group* e publicada em 2020 e sua terceira versão foi publicada em agosto de 2024.

Segundo o W3C (2024), o DCAT incorpora termos de vocabulários pré-existentes onde poderia-se identificar uma terminologia clara com significado apropriado, mas também inclui

definições informais resumidas externas para conveniência. As definições autorizadas são disponibilizadas nas referências normativas.

O DCAT foi utilizado no Catálogo Nacional de Dados como uma versão brasileira denominada DCAT-BR: “O DCAT-BR, especialização do DCAT, é um vocabulário do Brasil, um perfil usado para descrever conjuntos de dados do setor público e vinculados de forma padronizada” (Brasil, 2024a, p. 14). Segundo o órgão gestor, o DCAT-BR teve seu vocabulário e taxonomias adequados aos aspectos e legislações brasileiras, como a Lei Geral de Proteção de Dados Pessoais (LGPD). Anteriormente, o DCAT já havia ganhado especializações de outros governos, como o DCAT-AP (*Area of Policy*, para dados de interesse público), da Comissão Europeia e o DCAT-AP-NO, do governo da Noruega, para dados e serviços específicos do país (Brasil, 2024b).

Espera-se que, com o uso do vocabulário DCAT-BR, seja possível estabelecer uma consistência entre os agregadores de conteúdo e portais de conjuntos de dados por meio da padronização da descrição dos metadados dos conjuntos de dados. Acredita-se também que o vocabulário irá fornecer bases para a interoperabilidade necessária com outras interfaces. A especialização brasileira do vocabulário é mantida no âmbito da Política de Governança de Dados Nacional, pelo Grupo de Trabalho DCAT-BR, que é responsável pelas atualizações e alterações nas especificações do vocabulário (Brasil, 2024b).

4 PROCEDIMENTOS METODOLÓGICOS

Como mencionado, o universo desta pesquisa é o Catálogo Nacional de Dados (CND) do Governo Federal do Brasil, mantido pelo Ministério da Gestão e da Inovação em Serviços Públicos, em parceria com a Controladoria-Geral da União. O CND está disponível no endereço eletrônico <https://dados.gov.br/home>.

O acesso ao CND foi realizado entre os meses de janeiro e fevereiro do ano de 2025, com o objetivo de analisar empiricamente os dados e metadados utilizados, bem como a implementação do vocabulário semântico DCAT. Para embasar a análise também foram pesquisadas fontes documentais disponibilizadas pelos gestores do CND como materiais de

apoio⁷: Cartilha do Catálogo Nacional de Dados; Guia rápido do Catálogo Nacional de Dados e Catálogo Nacional de Dados: padrão DCAT-BR

Além disso, procedeu-se à pesquisa documental em fontes governamentais relacionadas à elaboração do CND, nos portais de dados abertos dos órgãos governamentais, bem como na documentação técnica do DCAT. Complementarmente, realizou-se uma pesquisa bibliográfica para fundamentação do referencial teórico. A análise da ferramenta considerou critérios relacionados à interoperabilidade, em consonância com as diretrizes estabelecidas na literatura especializada e nas normas vigentes no Brasil.

5 RESULTADOS

A partir da análise da documentação disponível sobre o CND, foram identificados os metadados utilizados para a descrição dos conjuntos de dados, denominados “campos” na Cartilha do CND (Brasil, 2024a), e os termos semânticos do vocabulário DCAT-BR, obtidos no Catálogo Nacional de Dados: padrão DCAT-BR (Brasil, 2024b). Com isso, foi realizado um mapeamento dos metadados por meio da sistematização e consolidação dessas informações em um quadro, considerando os termos do vocabulário semântico correspondentes em cada nível de descrição proposto pelo catálogo: bronze, prata e ouro. O resultado é apresentado no Quadro 1.

Quadro 1: Listas de Metadados e termos semântico do vocabulário DCAT-BR no Catálogo Nacional de Dados por nível de descrição

Bronze	Prata	Ouro
Organização (dcterms:Creator)	Área Geográfica (dcterms:spatial)	Substituído Por (dcterms:replaces)
Título (dcterms:title)	Data de Publicação (dcterms:issued)	Conjuntos de Dados Template
Descrição (dcterms:description)	Data de Alteração (dcterms:modified)	É Referenciado Por (dcterms:isReferencedBy)
Licença de uso (dcterms:license)	Idioma (dcterms:language)	Atualização de Versão (dcat:hasVersion)
Periodicidade de Atualização (dcterms:accrualPeriodicity)	Situação (adms:status)	URL de Obtenção de Acesso (dcat:landingPage)
Observância Legal (dcterms:accessRights)	Observância Legal Complementar (dcterms:rights)	Cobertura Temporal Fim (dcterms:temporal)

⁷ Disponível em: <https://www.gov.br/governodigital/pt-br/infraestrutura-nacional-de-dados/catalogo-nacional-de-dados/materiais-de-apoio>

Bronze	Prata	Ouro
Área Técnica Responsável (dcterms:Publisher)	Possui Relação com Objetivos de Desenvolvimento Sustentável (ODS)? (não possui termo semântico)	Cobertura Temporal Início (dcterms:temporal)
E-mail da Área Técnica Responsável (dcat:contactPoint)	Possui Dados de Gênero? (não possui termo semântico)	Cobertura Espacial
Palavras-Chave (dcat:keyword)	Possui Dados de Raça/Etnia? (não possui termo semântico)	Granularidade Espacial
Versão (dcat:version)	URL de Acesso (dcat:accessURL)*	Data da Descontinuação
Temas (dcat:theme)	Tipo de Mídia (dcat:mediaType)*	Conjuntos de Dados Associados (Interoperabilidade)
Dados abertos (não possui termo semântico)	Descrição do Recurso (dcterms:description)*	Tamanho em Bytes (dcat:byteSize)*
Visibilidade (não possui termo semântico)	Formato (dcterms:format)*	Observância Legal do Recurso (dcterms:accessRights)*
13	Título do Recurso (dcterms:title)*	Data de Publicação do Recurso (dcterms:issued)*
	14	Licença de Uso do Recursos (dcterms:license)*
		Data de Alteração do Recurso (dcterms:modified)*
		Verificação de Conteúdo (spdx:checksum)*
		17

Fonte: Baseado na Cartilha do CND e no Vocabulário Padrão DCAT-BR (os metadados de recursos foram indicados com asterisco).

Conforme descrito na Cartilha, há metadados para descrever recursos (em negrito no quadro 1) e metadados para descrever conjuntos de dados. A Cartilha não explica os termos, mas foram encontradas na página “Saiba como publicar um conjunto de dados”, do portal de Dados abertos do Gov.br, as seguintes definições: “Conjunto de Dados: Refere-se a uma coleção organizada de informações ou dados relacionados a um tema específico. [...] Recursos: São as representações individuais ou arquivos específicos que compõem um conjunto de dados” (Brasil, 2024a). Em termos simples, os recursos são itens específicos dentro de um conjunto de dados. Como continuidade da análise, os metadados foram avaliados quanto a sua utilidade para a descoberta de dados, ou seja, quais campos podem auxiliar o usuário a encontrar os dados de que necessita? O resultado da análise é apresentado no Quadro 2.

Quadro 2 - Metadados para descrição dos dados

Metadados para DESCRIÇÃO dos dados
1. Título (dcterms:title)
2. Descrição (dcterms:description)
3. Palavras-Chave (dcat:keyword)
4. Temas (dcat:theme)
5. Área Geográfica (dcterms:spatial)
6. Idioma (dcterms:language)
7. Possui Dados de Gênero?
8. Possui Dados de Raça/Etnia?
9. Descrição do Recurso (dcterms:description)
10. Título do Recurso (dcterms:title)
11. Cobertura Temporal Fim (dcterms:temporal)
12. Cobertura Temporal Início (dcterms:temporal)
13. Cobertura Espacial
14. Granularidade Espacial
15. Organização (dcterms:creator)
16. Periodicidade de Atualização (dcterms:accrualPeriodicity)
17. Data de Publicação (dcterms:issued)
18. Observância Legal Complementar (dcterms:rights)
19. URL de Obtenção de Acesso (dcat:landingPage)
20. Data de Alteração do Recurso (dcterms:modified)
21. Data de Alteração (dcterms:modified)
22. Conjuntos de Dados Associados (Interoperabilidade)

Fonte: Elaboração própria.

Dos 44 metadados propostos pelo CND, observou-se que 22 atendem ao objetivo de descoberta de dados. Os metadados que possuem maior potencial de contribuir para o objetivo de descoberta de dados são os metadados de descrição, pois tratam do conteúdo informacional dos dados descritos. Atributos como tema, palavras-chave, título, descrição, abrangência geográfica e temporal, entre outros, se referem ao conteúdo dos dados e fornecem bases para que o usuário encontre, identifique e selecione o item, ou seja, para a descoberta de dados.

Contudo, apesar da descoberta de dados ser um dos objetivos do CND e de que metade dos metadados propostos para catalogação dos dados sirvam a esse propósito, observa-se que o conteúdo mais relevante dos dados em si, que corresponde às variáveis contidas no conjunto de dados, não é obrigatoriamente registrado na catalogação do item. Por exemplo, se o usuário busca dados sobre "Taxa de participação na força de trabalho, por gênero, raça e faixa etária, de 2000 a 2024", é preciso que o conjunto de dados contenha as variáveis:

- personas na força de trabalho (isso pode não estar em um metadado);
- gênero;

c) raça e

d) faixa etária (que possa contemplar faixas pré-definidas em uma tabela).

As informações de gênero e raça estariam registradas nos metadados “Possui Dados de Gênero?” e “Possui Dados de Raça/Etnia?”, e o período dos dados nos metadados “Cobertura Temporal Início” e “Cobertura Temporal Fim”. Assim, os dados sobre as demais variáveis necessárias (força de trabalho, considerando pessoas em idade de trabalhar e faixa etária), não seriam descritas por nenhum metadado e mesmo que o catalogador utilizasse o `dcat:keywords`, seria necessário um mínimo de padronização nos registros para evitar níveis diferentes de especificidade da descrição.

Atualmente, a única forma de identificar quais dados um conjunto apresenta acontece apenas se o catalogador informar, durante a catalogação, no metadado “descrição” ou “descrição do recurso”, quais são essas variáveis. A outra forma de identificação só é possível após a obtenção dos dados. Porém, sendo o catálogo um instrumento destinado à identificação e seleção de itens, não deveria ser necessário que o usuário acessasse o conjunto de dados para confirmar se o conteúdo atende às suas necessidades.

Quanto à interoperabilidade, a documentação analisada indica que ela será proporcionada pelo uso do vocabulário semântico DCAT-BR, uma vez que este “fornece interoperabilidade semântica com outras interfaces de consumo de metadados com base na reutilização de vocabulários controlados estabelecidos e mapeamentos para vocabulários de metadados existentes.” (Brasil, 2024b, p. 6). Contudo, a partir do mapeamento dos metadados e termos do DCAT-BR apresentados no Quadro 1, verificou-se a existência de metadados sem termo semântico correspondente, bem como metadados diferentes que utilizam o mesmo termo semântico do DCAT-BR. Tal configuração pode comprometer a interoperabilidade semântica. Além disso, ressalta-se que essa interoperabilidade ocorre entre os metadados de representação dos dados do catálogo, e não entre os dados contidos nos conjuntos de dados e recursos catalogados.

6 CONSIDERAÇÕES FINAIS

O Catálogo Nacional de Dados atende ao que se propõe no que tange a servir como uma ferramenta para inventariar os dados produzidos pelo governo federal e incentivar a interoperabilidade semântica por meio da descrição dos dados dos diferentes órgãos de forma

padronizada, conforme disposto na literatura e normas disponíveis. No entanto, considerando-se os objetivos do CND como os de promover a descoberta de dados e a interoperabilidade semântica, identificou-se a possibilidade de aperfeiçoar a ferramenta.

Em relação ao objetivo da descoberta de dados, sugere-se a complementação da descrição dos dados com a inclusão de metadados que informem as variáveis que o recurso possui. Com isso, seria possível que o usuário soubesse se o conjunto de dados o atende antes mesmo de acessá-lo, cumprindo a função de um catálogo. O vocabulário DCAT não atenderia a essa necessidade, já que o mesmo não possui previsão de descrição de variáveis de dados em seus termos. No entanto, isso poderia ser incluído na especialização DCAT-BR.

Quanto ao objetivo de interoperabilidade, os termos do DCAT-BR não encontraram correspondência em todos os metadados para catalogação dos recursos e conjuntos de dados, o que pode prejudicar o processo. Observou-se também o uso de termos iguais para metadados diferentes, podendo ocasionar problemas na interoperabilidade.

A interoperabilidade semântica permite a colaboração e troca de dados e informações, independentemente de diferentes estruturas, processos de negócio e formas de operação em cada instituição, mas demanda que as organizações envolvidas conheçam seus processos de trabalho de forma a seguir padrões já utilizados. Caso contrário, será necessário trabalhar em outras soluções que contemplem a coleta e tratamento de dados e informações para o alinhamento destes, o que deverá ser realizado sempre que for necessário obter dados novos ou atualizados. Com a interoperabilidade semântica, esse trabalho à parte não é necessário, basta que seja feita uma única análise da compreensão necessária. Assim, padrões e processos já utilizados são mantidos, sendo necessário apenas definir os momentos para acionar as interações entre os sistemas.

REFERÊNCIAS

BRASIL. **Lei nº 12.527, de 18 de novembro de 2011**. Brasília, Presidência da República, 2011. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm. Acesso em: 13 jan. 2025.

BRASIL. Ministério da Gestão e Inovação em Serviços Públicos. Secretaria de Governo Digital. **Cartilha do catálogo nacional de dados**. Brasília: MGI, [2024a]. Disponível em: <https://www.gov.br/governodigital/pt-br/infraestrutura-nacional-de-dados/catalogo-nacional-de-dados/materiais-de-apoio/cartilha-do-catalogo-nacional-de-dados/@@download/file>. Acesso em: 13 jan. 2025.

BRASIL. Ministério da Gestão e Inovação em Serviços Públicos. Secretaria de Governo Digital. **Catálogo nacional de dados**: vocabulário padrão DCAT-BR. Brasília: MGI, [2024b]. Disponível em: <https://www.gov.br/governodigital/pt-br/infraestrutura-nacional-de-dados/catalogo-nacional-de-dados/materiais-de-apoio/vocabulario-padrao-dcat-br-v1.0/view>. Acesso em: 13 jan. 2025.

BRASIL. Ministério da Gestão e Inovação em Serviços Públicos. Secretaria de Governo Digital. **Guia rápido do Catálogo Nacional de Dados**. Brasília: MGI, [2024c]. Disponível em: <https://www.gov.br/governodigital/pt-br/infraestrutura-nacional-de-dados/catalogo-nacional-de-dados/materiais-de-apoio/guia-rapido-do-catalogo-nacional-de-dados>. Acesso em: 13 jan. 2025.

BRASIL. Ministério do Planejamento, Orçamento e Gestão. **Guia de interoperabilidade**: manual do gestor. Brasília: MP, 2012. Disponível em: <https://bibliotecadigital.economia.gov.br/handle/777/604>. Acesso em: 13 jan. 2025.

BRASIL. Portal Brasileiro de Dados Abertos. **Saiba como publicar um conjunto de dados**. [S. l.], [2025?]. Disponível em: <https://dados.gov.br/dados/conteudo/saiba-como-publicar-um-conjunto-de-dados>. Acesso em 22 fev. 2025.

CUNHA, M. B.; CAVALCANTI, C. R. de O. **Dicionário de Biblioteconomia e Arquivologia**. Brasília: Briquet de Lemos, 2008. 451 p.

IFLA. **Declaração dos Princípios Internacionais de Catalogação (PIC)**. 2018. Disponível em: <https://repository.ifla.org/items/195ded68-f032-4314-ac6e-a7ceafeef650>. Acesso em: 22 fev. 2025.

JAHNKE, N.; OTTO, B. Data catalogs in the enterprise: applications and integration. **Datenbank Spektrum**, n. 23, p. 89-96, jul. 2023. Disponível em: <https://doi.org/10.1007/s13222-023-00445-2>. Acesso em 22 fev. 2025.

WORLD WIDE WEB CONSORTIUM. **Data Catalog Vocabulary (DCAT)**: version 3. Disponível em: <https://www.w3.org/TR/vocab-dcat-3/>. Acesso em 13 jan. 2025.



[4.0 Internacional](https://creativecommons.org/licenses/by/4.0/)

Esta licença permite compartilhamento, remixe, adaptação e criação a partir do trabalho, mesmo para fins comerciais, desde que sejam atribuídos créditos ao(s) autor(es). Conteúdos elaborados por terceiros, citados e referenciados nesta obra não são cobertos pela licença.

Como citar este trabalho:

SOUZA, Katyusha Madureira Loures de; BRANDT, Mariana Baptista; VIDOTTI, Silvana Aparecida Borsetti Gregorio. *Catálogo Nacional de Dados do Governo Federal: análise da utilização do vocabulário DCAT-BR*. In: WORKSHOP DE INFORMAÇÃO, DADOS E TECNOLOGIA, 8., Marília, SP. **Anais [...]**. Marília, SP: Universidade de Marília. 2025. DOI: <https://doi.org/10.22477/viii.widat.253>.

Análise de dados para assessoramento legislativo: políticas públicas informadas por dados

Data analysis in legislative advising: data-informed public policies

Mariana Baptista Brandt¹



<http://lattes.cnpq.br/3761037263199030>



<https://orcid.org/0000-0001-8119-7527>

Resumo

Considerando a necessidade de legislações mais efetivas para a população brasileira e a crescente disponibilidade de dados e ferramentas tecnológicas para análise de grandes volumes, este artigo explora o uso da Análise de Dados e da Ciência de Dados como método para o assessoramento legislativo baseado em evidências. Para ilustrar essa abordagem, apresenta-se um estudo de caso que utiliza dados dos resultados das eleições municipais de 2024 na análise da política de cotas de raça e gênero, prevista em legislação eleitoral recente. O estudo seguiu as etapas de um processo de Análise de Dados e os resultados forneceram respostas quantitativas para as perguntas iniciais do projeto. Conclui-se que a Ciência de Dados pode ser uma ferramenta valiosa para o assessoramento legislativo, permitindo maior embasamento na formulação de políticas públicas.

Palavras-chave: análise de dados; políticas públicas informadas por dados; assessoramento legislativo; dados governamentais.

Abstract

Considering the need for more effective legislation for the Brazilian population and the increasing availability of data and technological tools for analyzing large volumes of information, this article explores the use of Data Analysis as a method to support evidence-based legislative advisory. To illustrate this approach, a case study is presented, utilizing electoral data to analyze race and gender quota policies in the 2024 municipal elections. The study follows all stages of the Data Analysis process, and the results provide quantitative answers to the project's initial questions. The findings suggest that Data Science can be a valuable tool for legislative advisory, enabling more informed and evidence-based policymaking.

Keywords: data analysis; data-informed public policies; legislative advising; government data.

¹ Doutora em Ciência da Informação, Universidade Estadual Paulista (Unesp), Marília, SP, Brasil. Analista legislativa, Câmara dos Deputados, Brasília, DF, Brasil. mariana.brandt@unesp.br.

1 INTRODUÇÃO

Os projetos de lei e demais propostas legislativas são feitas para provocar mudanças nas leis, buscando promover melhorias nos diversos setores da sociedade. Para isso, as casas legislativas contam com um quadro funcional especializado nas diversas áreas do conhecimento que prestam serviços de consultoria e assessoramento legislativo aos parlamentares. Volpe e Figueiredo Junior (2021) explicam que o trabalho das consultorias institucionais constitui uma base essencial para a atuação dos parlamentares, oferecendo apoio e orientação qualificada para que desempenhem suas funções de maneira dinâmica e eficaz, com compromisso técnico e isenção, atendendo a todas os partidos e correntes de pensamento.

Para que sejam efetivas, as mudanças devem ter em vista o que acontece, de fato, na sociedade brasileira. Um dos meios que pode ser utilizado para essa verificação é a análise de resultados de implementação das políticas públicas e de outros dados governamentais.

Assim, a análise de dados torna-se um importante instrumento de apoio ao assessoramento legislativo, viabilizando o chamado assessoramento legislativo baseado em evidências (ou informado por dados). Em pesquisa recente, Cambraia (2024) destaca que o assessoramento legislativo fundamentado em dados tem como propósito inovar a forma como as políticas públicas são elaboradas, debatidas e executadas no Brasil, assegurando maior alinhamento com as demandas reais da sociedade. O autor considera que diante de um cenário global cada vez mais dinâmico e interligado, torna-se evidente a importância de um processo legislativo mais célere, transparente e sustentado por dados concretos.

O objetivo deste trabalho é apresentar a Análise de Dados (*Data Analysis*) como método de obtenção de insumos informacionais para a elaboração legislativa por meio do assessoramento legislativo baseado em evidências, viabilizando a proposição de políticas públicas informadas por dados. Para isso, o artigo traz um estudo de caso de uso de dados de resultados eleitorais para subsidiar análises de assessoramento legislativo referente a política de cotas de raça e gênero nas eleições. O estudo justifica-se por seu valor social, na medida em que dá transparência para ações da prática legislativa, bem como por seu valor científico, apresentando estudo de caso de uso de Ciência de Dados, na prática.

A seguir, o referencial teórico apresenta conceitos chave do assessoramento legislativo, da análise de dados e de suas possibilidades de uso de dados para proposição, implementação e avaliação de políticas públicas.

2 ASSESSORAMENTO LEGISLATIVO

Em 2024 as consultorias legislativa e de orçamento da Câmara dos Deputados redigiram mais de 6 mil minutas de proposições², a pedido de deputados federais. Esse é um dos trabalhos que compõem a atuação dos órgãos no assessoramento legislativo institucional. Segundo Brasileiro (1968, p. 6), as estruturas de assessoramento legislativo pretendem “auxiliar o parlamentar a reunir objetivamente os dados de que tem necessidade, ajudá-lo a definir os problemas, a formular as alternativas de soluções, a pesar suas vantagens e inconvenientes e, finalmente, a estabelecer a opção mais aconselhável”. Este trabalho é previsto legalmente no capítulo V (Do Sistema de Consultoria e Assessoramento) do Regimento Interno da Câmara dos Deputados:

Art. 275. O sistema de consultoria e assessoramento institucional unificado da Câmara dos Deputados, além do Centro de Estudos e Debates Estratégicos, compreende a Consultoria Legislativa, com seus integrantes e respectivas atividades de consultoria e assessoramento técnico-legislativo e parlamentar à Mesa, às Comissões, às Lideranças, aos Deputados e à Administração da Casa, com o apoio dos sistemas de documentação e informação, de informática e processamento de dados. (Brasil, 2011, p. 158).

Para Horta (2011), o assessoramento legislativo deve atender a instituição legislativa para o cumprimento de suas funções: legiferante, parlamentar e fiscalizadora. Segundo o autor, o assessoramento legislativo ainda precisa evoluir. Essa evolução pode ser vista em consultorias de outros parlamentos, conforme aponta o estudo sobre a atuação de consultorias legislativas realizado por Vasconcelos *et al.* (2020), foi observado que:

Novas soluções para análise de dados – como big data e inteligência artificial – são utilizadas para o aperfeiçoamento das pesquisas e monitoramento das políticas públicas por vários parlamentos. Com isso parece ser possível disponibilizar maior quantidade e qualidade de informações, com o propósito de:

- a) facilitar o processo de tomada de decisão parlamentar;
- b) aperfeiçoar a prestação de contas à sociedade; e
- c) ampliar a capacidade de fiscalização do Poder Executivo, atribuição precípua do Parlamento. (Vasconcelos *et al.*, 2020, p. 25).

² Disponível em: <https://www2.camara.leg.br/transparencia/prestacao-de-contas/contas-da-camara/ano-de-2024/relatorio-de-gestao/rg-2024>.

Os autores identificaram que entre os países em que há áreas específicas para análise de dados nos parlamentos estão Japão, Chile, Estados Unidos e Reino Unido, além do parlamento europeu: “Verificou-se a implantação de setor especializado em análises estatísticas e ciência de dados, e ainda a constituição de parcerias, internas ou externas ao Parlamento, com órgãos especializados nos diversos temas.” (Vasconcelos *et al.*, 2020, p. 25).

No mesmo sentido, a Organização para Cooperação e Desenvolvimento Econômico (OCDE) aponta em relatório sobre governos guiados por dados que “o aumento dos dados e das tecnologias digitais está transformando economias e sociedades, com grandes implicações para as operações do dia a dia dos governos” (OCDE, 2019, p. 60, tradução nossa). O documento relata várias possibilidades de uso de dados pelos governos, incluindo a geração de valor para a sociedade por meio de atividades como:

- Antecipação e planejamento: uso de dados para projetos de políticas públicas, planejamento de intervenções, antecipação de possibilidades de mudança e previsão de necessidades
- Entregas: uso de dados para informar e melhorar implementação de políticas, responsividade dos governos e atividades de fornecimento de serviços públicos
- Avaliação e monitoramento: uso dos dados para mensurar impacto, auditar decisões e monitorar performance. (OCDE, 2019, p. 60, tradução nossa).

O uso de dados para projetos, implementação e avaliação de políticas públicas, como sugerido pela OCDE, demanda conhecimentos e práticas da Ciência de Dados, em especial da área de Análise de Dados, temas que serão tratados na seção a seguir.

3 CIÊNCIA DE DADOS, ANÁLISE DE DADOS E POLÍTICAS PÚBLICAS

A Análise de Dados (*Data Analysis*) pode ser considerada uma área ou disciplina da Ciência de Dados e consiste, de forma geral, no uso de dados para resolver problemas. Provost e Fawcett (2013) definem a Ciência de dados como um conjunto de princípios essenciais para obter conhecimento e *insights* a partir de dados de forma sistemática. Os autores afirmam que a “Ciência de dados envolve processos, princípios e ferramentas para compreender um fenômeno por meio da análise (automatizada) dos dados”. (Provost; Fawcett, 2013, p. 52-53, tradução nossa).

A Gartner (2025, tradução nossa) define *Data and Analytics* como “o gerenciamento de dados para todos os usos [...] e a análise de dados para conduzir processos de trabalho e melhorar resultados do negócio por meio de tomada de decisão mais efetiva [...]”. Há várias

definições para os termos relacionados a esse tema, porém, a discussão terminológica não é escopo desta pesquisa. Para fins deste artigo, utilizamos a definição de Análise de Dados como: “coleta, transformação e organização de dados com objetivo de endereçar conclusões, fazer previsões e direcionar decisões informadas” (Google, 202-?).

Anderson (2015) menciona seis abordagens de análise de dados, da mais simples à mais complexa, sendo elas as análises: descritiva, exploratória, inferencial, preditiva, causal e mecanicista, sendo essa última considerada pelo autor mais como um processo de modelagem do que como apenas um tipo de análise. Há outros autores que consideram quatro abordagens de análise: descritiva, diagnóstica, preditiva e prescritiva (Coursera, 2024). A análise descritiva usa a estatística descritiva como base para a análise, examinando os dados a partir das medidas de descrição estatística como média, mediana, desvio padrão e outras. É uma análise que informa “o quê” aconteceu. Já a análise diagnóstica identifica o “por quê” aconteceu o que foi identificado na análise descritiva. São análises que examinam o passado. Já as análises preditiva e prescritiva, olham para o futuro, indicando o que deve acontecer e o que fazer quanto ao que vai acontecer, respectivamente (Coursera, 2024).

São consideradas etapas fundamentais de um processo de Análise de Dados: identificar a questão que se quer responder, coletar dados que possam responder essa pergunta, limpar os dados para análise, analisar os dados utilizando técnicas e ferramentas, e interpretar os resultados da análise, verificando se a questão inicial foi respondida. (Coursera, 2024). Há variações dessas fases, como: fazer perguntas, preparar os dados, processar os dados (limpeza), analisar os dados limpos, compartilhar os achados e recomendações e agir com base nos resultados (Google, 202-?). Contudo, as fases propostas pelos diferentes autores podem ser consideradas equivalentes.

A Ciência de Dados, embora inicialmente desenvolvida no contexto empresarial, também pode ser aplicada no setor público para aprimorar a tomada de decisões. Seu uso possibilita uma compreensão mais profunda e transparente da realidade, além de melhorar a identificação e seleção de alternativas em processos decisórios (Arnaboldi; Azzone, 2020). Assim, os dados possuem um papel fundamental:

Os dados são necessários para execução de serviços públicos. Para elaborar políticas públicas que atendem às reais necessidades da sociedade, especialmente para conhecer o público-alvo: suas características demográficas, idade, sexo, geográficas, de renda, entre outras. (Brasil, 2024, p. 8).

Reconhecendo a importância da produção e uso de dados no setor público, o governo brasileiro, por meio da Controladoria-Geral da União (CGU), lançou um Plano Nacional de Governo Aberto, que estabelece compromissos voltados à produção de dados pelos órgãos. Entre eles, destaca-se o compromisso 5, relacionado a dados de ações afirmativas étnico-raciais. O compromisso objetiva “Produzir, tratar e disponibilizar dados de ações afirmativas étnico-raciais padronizados e de qualidade, de forma colaborativa e em catálogo único.” (Brasil, 2023, p. 35). A produção de dados étnico-raciais, bem como dados de gênero, é essencial para avaliação das políticas públicas, pois esses recortes impactam na aplicação da política pública. A análise dos dados deve ser estratificada para garantir maior equidade e efetividade nas decisões governamentais, já que as políticas públicas não afetam todas as pessoas da mesma maneira.

Steif (2021), na introdução de seu livro *Public policy analytics*, afirma que o mais comum é que as políticas públicas orientadas por evidências utilizem dados para tomada de decisão orçamentária, mas a tomada de decisão operacional sobre a implementação da política é igualmente importante. Em comparação com empresas, que possuem como indicador claro de sucesso o lucro, nos governos há várias métricas de sucesso. Portanto devem-se analisar as diversas dimensões: econômicas, políticas, burocráticas e de equidade, entre outras. (Steif, 2021).

Nesse sentido, Arnaboldi e Azzone (2020) realizaram um estudo com o objetivo de investigar como a Análise de dados e a Ciência de dados podem fornecer subsídios informacionais para a elaboração de políticas públicas. As autoras afirmam que encontraram muitos estudos anteriores sobre o uso de Ciência de dados no setor público, mas a maioria não tratava do resultado ou da decisão tomada, ou seja, não relatam se a informação resultante do processo foi útil ou não na elaboração de políticas públicas. A partir da análise de 3 estudos de caso, as autoras chegaram a algumas conclusões. A primeira foi de que o papel-chave nesse processo é o de um tradutor, ou seja, uma pessoa que traduza o trabalho de ciência de dados para quem propõe as políticas públicas e vice-versa. Outro resultado observado foi que a fase de aquisição dos dados é particularmente crítica e demanda uma função dedicada para isso. Além disso, as autoras relatam que a visualização de dados é essencial para construir confiança no tradutor e o valor da visualização aumenta conforme o escopo dos dados. Por fim, concluem que o resultado do processo é influenciado pela

qualidade dos dados, pela clareza das opções para as políticas públicas e pela forma como os dados contribuem no sentido de simular efeitos legislativos e financeiros.

4 PROCEDIMENTOS METODOLÓGICOS

Trata-se de uma pesquisa aplicada, com estudo de caso do uso de análise de dados no contexto dos resultados de eleições municipais de 2024, com o objetivo de subsidiar o assessoramento legislativo na área de Direito Eleitoral na Câmara dos Deputados. Além do estudo de caso, foi realizada uma entrevista com os consultores responsáveis pela demanda de dados para averiguar se os dados fornecidos foram relevantes para o trabalho de assessoramento legislativo.

4.1 Estudo de Caso

O trabalho teve início durante o período de eleições municipais de 2024, quando, semanas antes do primeiro turno, surgiram alguns questionamentos referentes ao impacto que uma mudança recente na legislação eleitoral poderia causar nas eleições subsequentes. As alterações na lei foram referentes a novas regras para cotas de gênero e de raça para as candidaturas e quanto à distribuição dos fundos eleitoral e partidário, também com base nos critérios de raça e gênero. Para isso, foi elaborada uma série de perguntas que deveriam ser respondidas com dados dos resultados das eleições municipais.

Assim, o processo de análise de dados teve início com uma análise inicial das perguntas feitas pelos consultores, seguida do estudo da estrutura dos dados de resultados eleitorais disponibilizados em sua fonte original, o Tribunal Superior Eleitoral (TSE). O órgão disponibiliza um portal de dados abertos³ com mais de 160 conjuntos de dados referentes a candidaturas, apurações, resultados, prestação de contas, entre outros. Verificou-se que essas perguntas poderiam ser respondidas com os dados dos conjuntos de dados referentes a candidaturas, que possuem dados sobre os candidatos, incluindo gênero, raça/cor, partido, estado, situação da candidatura, escolaridade, dentre outros. Ao final de cada turno, os dados foram atualizados com a inclusão dos resultados das eleições, por meio de uma variável que informa

³ Disponível em: <https://dadosabertos.tse.jus.br/>.

a situação do candidato como “eleito; não eleito”. Assim, foram utilizados os conjuntos de dados das eleições de 2020 e 2024, pois o objetivo foi a comparação entre os resultados desses pleitos, considerando as mudanças na legislação eleitoral.

Após a obtenção dos dados, que foi realizada a partir do download de arquivos no formato csv, os mesmos foram inseridos no PowerBI, software para tratamento e visualização de dados. A seguir, foi realizada uma verificação para identificar se havia necessidades de limpeza dos dados. Não foram encontradas duplicações, inconsistências ou problemas de padronização, estrutura ou formato. Também não foram encontrados problemas quanto à completude, atualização ou confiabilidade, o que se acredita que ocorreu devido à própria natureza dos conjuntos de dados utilizados: dados que haviam sido produzidos recentemente, são de grande interesse público e possuem órgão responsável com estrutura administrativa permanente responsável pela gestão do serviço de Dados Abertos. Considerou-se então que os dados estavam prontos para uso, não havendo necessidade de tratamento.

A próxima etapa consistiu em uma análise mais profunda das perguntas realizadas pelos consultores legislativos, às quais deveriam ser respondidas pelos dados obtidos após o resultado das eleições de 2024. A partir das perguntas, foram construídos indicadores e a partir deles, as variáveis necessárias para responder a cada pergunta foram identificadas no conjunto de dados coletado, conforme exemplificado no Quadro 1.

Quadro 1 - Exemplos de perguntas, indicadores e variáveis

Perguntas	Indicadores	Variáveis utilizadas
Quantas mulheres foram eleitas em municípios para o cargo de prefeita e vice-prefeita?	Número de mulheres eleitas para prefeita; Número de mulheres eleitas para vice-prefeita	Gênero= feminino; Cargo= vereador; prefeito; Resultado= eleito
Quantos candidatos negros (pretos e pardos – autodeclaração) foram eleitos prefeitos e nas CV?	Número de pessoas negras eleitas para prefeito; Número de pessoas negras eleitas para vereador	Resultado= eleito; Cor/raça= Preta, Parda; Cargo=vereador, prefeito
Quantas foram as Câmaras com apenas 1 mulher?	Número de Câmara Municipais com apenas 1 mulher eleita	Gênero= feminino; Cargo= vereador; Resultado = eleito; Nome do município

Fonte: Elaboração própria.

O TSE disponibiliza dicionários de dados explicando o significado das variáveis, o que foi de grande auxílio na identificação das informações necessárias. A demanda inicial dos consultores constou de 15 perguntas, das quais 13 poderiam ser respondidas com dados. As

medidas necessárias foram construídas no PowerBI por meio da utilização da linguagem DAX, própria da ferramenta. Foram criadas 58 medidas, que são dados derivados da agregação dos dados brutos coletados do TSE. Com isso, foi elaborado um painel de dados que pudesse trazer, em dados, as respostas necessárias.

4.2 Entrevista com Consultores Legislativos

Como parte da pesquisa para este artigo, foi realizada uma entrevista com os consultores legislativos para averiguar a efetividade do produto de dados produzido no âmbito de seu trabalho de assessoramento legislativo. A entrevista com um dos consultores foi realizada em 21/02/2025, de forma presencial, em Brasília, e foi do tipo semiestruturada, com 4 perguntas previamente preparadas para guiar a coleta de dados. Outro consultor respondeu a pesquisa por e-mail, com as mesmas 4 perguntas, em 24/02/2025.

Os resultados e discussão do estudo de caso serão apresentados na seção a seguir.

5 RESULTADOS

O estudo de caso permitiu observar a aplicação das etapas do processo de análise de dados mencionadas na seção 3. O Quadro 2, a seguir, traz essa correspondência.

Quadro 2 - Etapas do processo de análise de dados

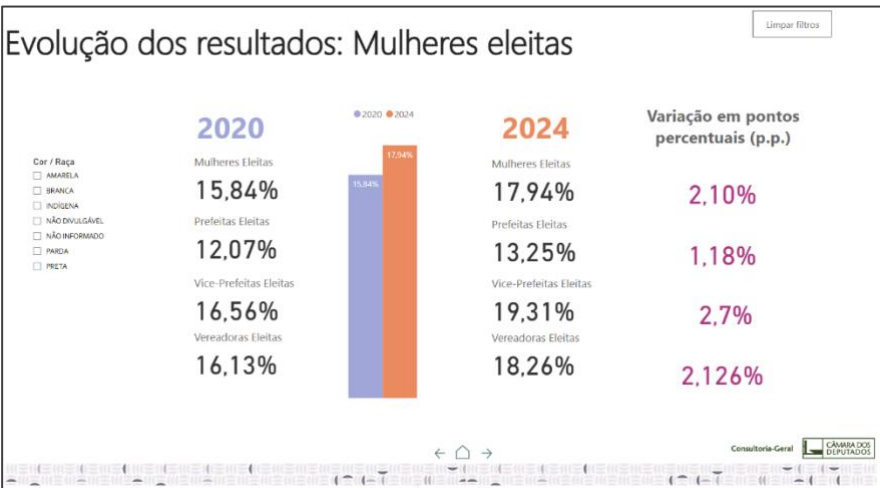
Etapa	Estudo de caso “Eleições Municipais 2024”
1) Identificar a questão que se quer responder	Os consultores legislativos elaboraram 15 perguntas a respeito das mudanças na legislação eleitoral, das quais verificou-se que 13 poderiam ser respondidas com dados.
2) Coletar dados que possam responder essa pergunta	Os dados foram coletados no portal de Dados Abertos do Tribunal Superior Eleitoral, que é a fonte primária para dados de resultados eleitorais brasileiros.
3) Limpar os dados para análise	Não houve necessidade visto que os dados possuíam a qualidade necessária para a análise.
4) Analisar os dados utilizando técnicas e ferramentas	Os dados foram inseridos na ferramenta de análise PowerBI, onde foram elaboradas 58 medidas necessárias para responder às perguntas iniciais. Foram elaborados 10 painéis com visualizações de dados.

5) Interpretar os resultados da análise, verificando se a questão inicial foi respondida	Os consultores receberam e analisaram o painel de dados e, em entrevista, responderam que consideram que as questões foram plenamente respondidas a partir dos dados entregues.
--	---

Fonte: Elaboração própria.

Ao final, foram construídas 10 páginas para fornecer respostas às 13 perguntas iniciais dos consultores legislativos, os quais receberam e validaram o painel como um todo. A implementação dos gráficos e demais visuais do painel foram feitas com base nas melhores práticas de visualização de dados, arquitetura da informação e design visual.

Figura 1 - Recorte do painel de dados de eleições municipais



Um resumo das métricas e dimensões de análise implementadas no painel de dados está representado no Quadro 3.

Quadro 3 - Métricas e dimensões de análise

	Número de candidatos	Número de eleitos	Percentual de candidatos	Percentual de eleitos
Sexo	X	X	X	X
Cor/Raça	X	X	X	X
Partido	X	X		
Ano eleição	X	X	X	X
Cargo	X	X	X	X
UF	X	X		
Município	X	X		

Fonte: Dados da pesquisa (2024).

⁴ Disponível em: <https://cd.leg.br/dados-eleicoes-municipais>.

A motivação para a demanda de análise dos dados de resultados das eleições municipais de 2024 surgiu, segundo os consultores entrevistados, da necessidade de analisar o impacto das alterações recentes nas regras eleitorais em relação a candidaturas e distribuição dos fundos eleitoral e partidário, bem como mensurar a efetividade da ação afirmativa (cotas de gênero e raça). Além disso, estudos e simulações sobre dados reais de eleições passadas com o propósito de antecipar possíveis efeitos decorrentes da alteração legislativa são importantes para uma melhor projeção legislativa, relatam os consultores. Sobre o produto apresentado como resultado desse estudo de caso, ou seja, o painel de dados de eleições municipais, um dos consultores entrevistados considerou que o painel mostra claramente a sub-representação feminina: somente 17,94% dos eleitos em 2024 são mulheres, com um crescimento de apenas 2,1 pontos percentuais em relação a 2020. Além disso, o uso de dados permite subsidiar a decisão política do parlamentar com maior segurança. O segundo consultor informou que considera “o painel de dados um produto excelente. Permite a análise de informações fidedignas oriundas das bases de dados da Justiça Eleitoral e que permite realizar cruzamentos de dados de interesse dos parlamentares. Essas informações são cruciais para subsidiar decisões políticas legislativas.”

Os consultores consideram que os dados para o assessoramento legislativo são úteis tanto na avaliação das políticas públicas atuais quanto na proposição de propostas de aperfeiçoamento das mesmas e que, sem o assessoramento com dados, as propostas legislativas seriam como um “voo cego”. Em relação às possibilidades de contribuição da análise de dados para o assessoramento legislativo, os consultores entendem ser uma prática essencial para as grandes democracias a que a análise de dados é indispensável atualmente, além de auxiliar no convencimento dos deputados: “o uso de dados é um argumento muito forte para a discussão com os parlamentares”. O consultor afirma que o assessoramento legislativo baseado em dados já é uma realidade em diversos parlamentos ao redor do mundo e que, no Brasil, essa prática ainda está em estágio inicial, mas tende a se consolidar à medida que as tecnologias avançam e o acesso a dados, bem como sua análise em tempo hábil, se tornam mais viáveis. Por fim, o consultor afirma que propor uma nova legislação “sem os dados, sem conhecer a realidade, é um tiro no escuro”. A partir das respostas obtidas na entrevista, considera-se que o produto de dados entregue foi efetivo para o trabalho dos consultores legislativos.

5 CONSIDERAÇÕES FINAIS

A Análise de dados para o assessoramento legislativo é realidade em vários parlamentos e a Ciência de Dados vem a ser uma ferramenta cada vez mais utilizada no âmbito da elaboração de políticas públicas. Os órgãos de assessoramento como consultorias e assessorias que possuem estrutura para viabilizar o fornecimento de dados para os consultores de forma sistematizada, com base nas etapas, métodos, técnicas e ferramentas da Ciência de Dados e da Análise de Dados e de forma ágil, poderão fornecer um assessoramento técnico mais assertivo e embasado, informado por dados, elevando a qualidade das propostas legislativas futuras.

O estudo de caso trouxe um exemplo bem-sucedido do uso de dados para avaliação do impacto de uma mudança na legislação eleitoral brasileira, com uso de dados da Justiça Eleitoral. O processo realizado seguiu as etapas formais consideradas na Análise de dados: identificar a questão, coletar dados, limpar os dados (se necessário), analisar os dados e interpretar os resultados, sendo esta última etapa realizada por consultores legislativos que assessoram diretamente os deputados federais. O produto resultante da análise de dados foi um painel de dados com 10 páginas que respondeu a 13 questões específicas elaboradas pelos consultores legislativos.

Retomando o histórico do assessoramento legislativo trazido na seção 2, Horta (2011, p. 46) considera-se que “[...] o assessoramento legislativo na Câmara dos Deputados, ainda não chegou ao fim de sua trajetória evolutiva. Tem de lutar ainda contra carências, distorções e incompreensões para alcançar o nível ótimo de estruturação e desempenho”. Acredita-se que a aplicação de métodos e técnicas de Análise de Dados e Ciência de Dados seja um caminho para impulsionar a evolução do assessoramento legislativo para uma nova fase.

REFERÊNCIAS

ANDERSON, C. **Creating a data-driven organization**. Sebastopol, CA: O’Reilly Media, 2015.

ARNABOLDI, M; AZZONE, G. Data science in the design of public policies: dispelling the obscurity in matching policy demand and data offer. **Heliyon**, v. 6, n. 6, 2020. Disponível em: [https://www.cell.com/heliyon/fulltext/S2405-8440\(20\)31144-0?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS2405844020311440%3Fshowall%3Dtrue](https://www.cell.com/heliyon/fulltext/S2405-8440(20)31144-0?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS2405844020311440%3Fshowall%3Dtrue). Acesso em: 17 jan. 2025.

BRASIL. Congresso Nacional. Câmara dos Deputados. **Regimento interno da Câmara dos Deputados**. 9. ed. Brasília: Câmara dos Deputados, Edições Câmara, 2011.

BRASIL. Controladoria-Geral da União. **6º Plano de ação nacional em governo aberto: open government partnership 2023/2027**. 2023. Disponível em: https://www.gov.br/cgu/pt-br/governo-aberto/a-ogp/planos-de-acao/6deg-plano-de-acao-brasileiro/brazil_action-plan_2023-2027_december_pt.pdf. Acesso em: 24 fev. 2025.

BRASIL. Ministério da Gestão e Inovação em Serviços Públicos. Secretaria de Governo Digital. **Cartilha do Catálogo Nacional de Dados**. 2024. Disponível em: https://www.gov.br/cgu/pt-br/governo-aberto/a-ogp/planos-de-acao/6deg-plano-de-acao-brasileiro/brazil_action-plan_2023-2027_december_pt.pdf. Acesso em: 24 fev. 2025.

BRASILEIRO, A. M. **O assessoramento legislativo**. Rio de Janeiro: FGV, 1968.

CAMBRAIA, T. Assessoramento legislativo baseado em dados: análise da mulher no mercado de trabalho. **Agenda brasileira**, v. 5, n. 9, 2024. Disponível em: <https://bd.camara.leg.br/bd/handle/bdcamara/42041>. Acesso em 07 jan. 2025.

COURSERA. **What Is data analysis?** 2024. Disponível em: <https://www.coursera.org/articles/what-is-data-analysis-with-examples>. Acesso em: 15 jan. 2025.

GARTNER. **Information technology glossary**. 2025. Disponível em: <https://www.gartner.com/en/information-technology/glossary/data-and-analytics>. Acesso em: 19 fev. 2025.

GOOGLE. **Foundations: data, data, everywhere**. Curso on-line. [2025]. Disponível em: <https://www.coursera.org/learn/foundations-data/>. Acesso em: 17 jan. 2025.

HORTA, A. Breve memória sobre o assessoramento legislativo na Câmara dos Deputados. *In: 40 anos de Consultoria Legislativa: consultores legislativos e consultores de orçamento*. Brasília: Câmara dos Deputados, Edições Câmara, 2011. Disponível em: <http://bd.camara.leg.br/bd/handle/bdcamara/7465>. Acesso em: 15 jan. 2025.

ORGANIZAÇÃO PARA COOPERAÇÃO E DESENVOLVIMENTO ECONÔMICO (OCDE). **The path to becoming a data-driven Public Sector**. Paris: OECD Digital Government Studies, OECD Publishing, 2019. Disponível em: https://www.oecd.org/en/publications/the-path-to-becoming-a-data-driven-public-sector_059814a7-en.html. Acesso em: 15 jan. 2025.

PROVOST F.; FAWCETT, T. Data science and its relationship to big data and data-driven decision making. **Big Data**, v. 1, n. 1, p. 51-59, 2013. DOI: <https://doi.org/10.1089/big.2013.1508>. Disponível em: <https://www.liebertpub.com/doi/epdf/10.1089/big.2013.1508>. Acesso em: 15 jan. 2025.

STEIF, K. **Public policy analytics: code and context for data science in government**. Boca Raton: CRC Press, 2021. Disponível em:

<https://www.taylorfrancis.com/books/mono/10.1201/9781003054658/public-policy-analytics-ken-steif>. Acesso em: 15 jan. 2025.

VASCONCELOS, B., T. *et al.* **Análise exploratória de consultorias legislativas**. Brasília: Câmara dos Deputados, 2020. Disponível em:

<https://bd.camara.leg.br/bd/handle/bdcamara/40708>. Acesso em 07 jan. 2025.

VOLPE, R.; FIGUEIREDO JUNIOR, W. Futuro das consultorias institucionais do poder legislativo. *In*: PINHEIRO, A.; NOGUEIRA, C. (org.) **História viva da Consultoria Legislativa**: edição comemorativa de 50 anos. Brasília: Câmara dos Deputados, Edições Câmara, 2021. Disponível em: <https://bd.camara.leg.br/bd/handle/bdcamara/40659>. Acesso em: 06 jan. 2025.



[4.0 Internacional](https://creativecommons.org/licenses/by/4.0/)

Esta licença permite compartilhamento, remixe, adaptação e criação a partir do trabalho, mesmo para fins comerciais, desde que sejam atribuídos créditos ao(s) autor(es). Conteúdos elaborados por terceiros, citados e referenciados nesta obra não são cobertos pela licença.

Como citar este trabalho:

BRANDT, Mariana Baptista. *Análise de dados para assessoramento legislativo: políticas públicas informadas por dados*. *In*: WORKSHOP DE INFORMAÇÃO, DADOS E TECNOLOGIA, 8., Marília, SP. **Anais [...]**. Marília, SP: Universidade de Marília, 2025. DOI: <https://doi.org/10.22477/viii.widat.235>.



WIDaT 2025

VIII WORKSHOP DE INFORMAÇÃO,
DADOS E TECNOLOGIA

MARÍLIA • SP

Ciência de Dados e Inteligência Artificial no Poder Judiciário brasileiro

Data Science and Artificial Intelligence in the Brazilian Judiciary

Fabício Fracaroli Pereira¹



<http://lattes.cnpq.br/0730277352664783>



<https://orcid.org/0009-0003-6959-8557>

Deise Maria Antonio Sabbag²



<http://lattes.cnpq.br/0772836405405573>



<https://orcid.org/0000-0001-6392-4719>

Resumo

O artigo tem como objetivo analisar a aplicação da Ciência de Dados e da Inteligência Artificial (IA) no âmbito jurídico, com foco no Poder Judiciário brasileiro. A pesquisa se baseia em dados do Conselho Nacional de Justiça (CNJ), que avalia anualmente o uso de IA no Judiciário desde 2020. A IA, com sua rápida evolução, e o Big Data, que busca extrair informações valiosas de grandes volumes de dados, têm atraído atenção significativa por suas potencialidades. O tema é relevante e atual, e o estudo justifica-se também pelo interesse do autor na aplicação dessas tecnologias no setor jurídico. A pesquisa inclui uma revisão bibliográfica sobre Ciência de Dados e IA analisando o relatório mais recente do CNJ sobre o uso de IA no Judiciário. Os resultados mostram que 66% dos tribunais que responderam à pesquisa têm projetos de IA em desenvolvimento, com os Tribunais Superiores apresentando a maior média de projetos. Os principais benefícios identificados são maior eficiência e agilidade no processamento de documentos. Além disso, há uma cooperação significativa com a iniciativa privada e universidades. No entanto, existem preocupações com riscos técnicos e questões éticas que podem afetar a implementação e desenvolvimento desses projetos.

Palavras-chave: Ciência de Dados; Inteligência Artificial; Poder Judiciário.

Abstract

The article aims to analyze the application of Data Science and Artificial Intelligence (AI) in the legal field, with a focus on the Brazilian Judiciary. The research is based on data from the National Justice Council (CNJ), which annually evaluates the use of AI in the Judiciary since 2020. AI, with its rapid evolution, and Big Data, which seeks to extract valuable information from large volumes of data, have attracted significant attention due to their potential. The topic is relevant and current, and the study is also justified by the author's interest in the application of these technologies in the legal sector. The research includes a literature review on Data Science and AI, analyzing the most recent CNJ report on the use of AI in the Judiciary. The results show that 66% of the courts that responded to the survey have AI projects in development, with the Superior Courts having the highest average of projects. The main benefits identified are increased efficiency

¹ Doutorando pelo Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual Paulista (Unesp), Marília, São Paulo, Brasil. Bolsista CAPES PROEX. fabriciofracaroli@gmail.com.

² Doutora em Ciência da Informação pela Universidade Estadual Paulista (Unesp), Marília, São Paulo, Brasil. Docente da Universidade de São Paulo (USP), Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, São Paulo, Brasil. deisesabbag@usp.br.

and agility in document processing. Furthermore, there is significant cooperation with private companies and universities. However, there are concerns regarding technical risks and ethical issues that may affect the implementation and development of these projects. 1616

Keywords: Data Science; Artificial Intelligence; Judiciary.

1 INTRODUÇÃO

O presente artigo buscou avaliar o panorama atual sobre o uso de mecanismos de Inteligência Artificial (IA) no Poder Judiciário brasileiro, mediante análise de pesquisa realizada pelo Conselho Nacional de Justiça (CNJ) que objetivou “[...] acompanhar o desenvolvimento de soluções de IA no Judiciário brasileiro” (CNJ, 2024, p. 13).

A Inteligência Artificial tem atraído olhares atentos nos últimos anos, especialmente em razão da sua surpreendente evolução e diante das inúmeras possibilidades que representa para o aprimoramento das atividades que dela se utilizam. Da mesma forma, o processamento de uma quantia gigantesca de dados – *Big Data* – coletados nas mais diversas interações humanas e que, com recurso à Ciência de Dados, busca extrair valiosas informações desse conjunto, tem recebido especial atenção.

Assim, a escolha do tema justifica-se pela sua atualidade e importância, e o foco no Poder Judiciário tem fundamento na formação básica em nível superior do autor (o Direito), bem como pelo seu interesse no uso da Inteligência Artificial e da Ciência de Dados para a realização de atividades voltadas a esse ramo do conhecimento. Do mesmo modo, pela interdisciplinaridade intrínseca do tema, optou-se por estudar a atual situação dos órgãos judiciais nacionais.

A pesquisa também pode ser classificada como um estudo de caso³, realizada mediante revisão bibliográfica sobre os temas *Ciência de Dados e Inteligência Artificial*, bem como assuntos correlatos, a fim de contextualizar sobre o seu objeto para, em seguida, proceder a uma análise da *Pesquisa uso de Inteligência Artificial (IA) no Poder Judiciário 2023*, estudo que é elaborado anualmente pelo Conselho Nacional de Justiça desde o ano de 2020 com o objetivo de investigar a situação atualmente existente nos

³ “A metodologia do Estudo de Caso é uma abordagem qualitativa do objeto, muito utilizada nas ciências sociais aplicadas, tendo em Robert K. Yin (2015) um dos principais teóricos. O estudo de caso é definido como uma pesquisa empírica que investiga um fenômeno contemporâneo dentro do seu contexto real de ocorrência, sendo que este fenômeno é estudado em profundidade e de forma holística, com o objetivo de compreender a complexidade e a dinâmica do objeto em questão” (Assis; Monteiro, 2023, p. 14).

órgãos mencionados sobre a temática abordada, já que esta é a mais recente pesquisa publicada pelo CNJ sobre o tema.

Optou-se pela análise da pesquisa elaborada pelo CNJ por se tratar da publicação mais recente disponível (atualidade), realizada anualmente (regularidade), e proveniente de órgão cuja principal atribuição consiste no controle da atuação administrativa e financeira do Poder Judiciário e do cumprimento dos deveres funcionais dos juízes, nos termos do artigo 103-B, §4º, da Constituição Federal brasileira.

A abordagem metodológica adotada é predominantemente quantitativa, com foco na análise do número de tribunais e conselhos que informaram possuir projetos de Inteligência Artificial e da quantidade de projetos por órgão. Complementarmente, também foram consideradas análises qualitativas, que contribuíram para a interpretação dos dados e fundamentação das conclusões do estudo.

2 CIÊNCIA DE DADOS E INTELIGÊNCIA ARTIFICIAL

A Inteligência Artificial certamente tem sido um dos temas com maior repercussão nos últimos tempos. Embora não seja algo tão recente – seus primórdios remontam à década de 1950⁴ –, nos últimos anos a evolução dessa tecnologia passou a, cada vez mais, fazer parte do cotidiano das pessoas, com ou sem seu conhecimento – e consentimento. Não raro, nos deparamos com informações sobre o uso de IA em sistemas de busca na *web*, em aplicativos de mobilidade e no uso de *chatbots* em interações com empresas, ou mesmo em serviços de *streaming*, por exemplo. E, também, muitas vezes somos surpreendidos com sua utilização em situações que não cogitamos, como acontece em *sites* e ferramentas de *e-commerce* que nos indicam – ou induzem – a compra de determinados produtos ou mesmo em exames e diagnósticos relacionados à saúde.

Por *Inteligência Artificial* podemos compreender o campo do conhecimento ligado a características predominantemente humanas e suas interações com máquinas,

⁴ “Na década de 1950, no âmbito dos esforços científico-tecnológicos de criar modelos de simulação da mente humana, surgiu o campo da inteligência artificial” (Kaufman, 2019, p. 21). Sobre os primórdios da inteligência artificial, Boden (2020) esclarece que já na década de 1840 a matemática inglesa Ada Lovelace antecipou o que hoje pode ser considerada a inteligência artificial (ou, ao menos, parte dela), por exemplo, ao afirmar que uma máquina analítica poderia compor peças musicais de qualquer nível de complexidade, bem como por descrever princípios de programação modernos.

como linguagem, inteligência, raciocínio, aprendizagem e resolução de problemas, de modo que possibilite uma simbiose entre o humano e a máquina, seja por intermédio de dispositivos acoplados ao corpo humano ou mesmo mediante uma interatividade ou conexão intelectual (Kaufman, 2019). Desde dispositivos inteligentes, como celulares (*smartphones*), luminárias (*smart lamps*) e caixas de som (*smart speakers*), como em acessórios vestíveis (*wearables*), a Inteligência Artificial desempenha importante papel na análise e processamento dos dados coletados. E aqui se encontra um conceito de notável protagonismo: o **dado**.

Para Setzer (2015, p. 1), dado é uma “(...) sequência de símbolos quantificados ou quantificáveis”:

Quantificável significa que algo pode ser quantificado e depois reproduzido sem que se perceba a diferença para com o original. Portanto, um texto é um dado. De fato, as letras são símbolos quantificados, já que o alfabeto, sendo um conjunto finito, pode por si só constituir uma base numérica (a base hexadecimal empregada em geral nos computadores usa, além dos 10 dígitos decimais, as letras de A a E). Também são dados fotos, figuras, sons gravados e animação, pois todos podem ser quantificados ao serem introduzidos em um computador, a ponto de se ter eventualmente dificuldade de distinguir a sua reprodução com o original. É muito importante notar-se que, mesmo se incompreensível para o leitor, qualquer texto constitui um dado ou uma sequência de dados (Setzer, 2015, p. 1).

No contexto das tecnologias digitais, praticamente toda interação gera uma quantia considerável de dados, intencionalmente ou não. Dados que são capturados, tratados e muito bem utilizados pelas empresas. Como ensina Kaufman (2019, p. 33):

Qualquer interação com tecnologias digitais deixa "rastros", alguns voluntários como as publicações nas redes sociais - Facebook, Twitter e Instagram -, e outros involuntários, como as informações armazenadas nos bancos de dados digitais na compra com cartão de crédito, na movimentação bancária on-line, no acesso aos programas de fidelidade, no vale-transporte, nas comunicações por telefonia móvel, no acesso online à resultados de exames médicos, e inúmeras outras ações presentes em nossa rotina. Esses podem ser usados pelas plataformas originais ou "reusados" por terceiros, ou combinados pela fusão de conjunto de dados, com as mais variadas finalidades, e são responsáveis por inúmeros benefícios da sociedade do século XXI (e, igualmente, por inúmeras ameaças).

Os *rastros* deixados pelos indivíduos no mundo virtual (e, também, fisicamente), por mais irrelevantes que possam parecer, são dados importantíssimos. Isoladamente, podem não dizer muito; mas quando processados e analisados em conjuntos com outros

dados, objetivo precípuo da Ciência de Dados, são fontes de preciosas informações. Com a geração cada vez mais intensa de dados pelos sujeitos que se utilizam dos serviços *on-line*, seus **controladores**⁵ encontram uma fonte quase inesgotável de possibilidades, no sentido de que, com um tratamento adequado, é possível alcançar inúmeras informações sobre os usuários – os **titulares**⁶ de tais dados –, isolada ou conjuntamente, que podem ser convertidas em retornos financeiros altamente rentáveis.

E nesse contexto também é possível identificar outro importante fenômeno, que diz respeito ao chamado *Big Data*, expressão relativa ao “[...] termo em inglês que descreve o grande volume de dados gerados e armazenados [...]” (Kaufman, 2019, p. 32), frequentemente utilizados para abastecer sistemas de análise de dados em grande escala, além de serem utilizados para treinar sistemas de Inteligência Artificial.

O conceito de *Big Data* há muitos anos tem ganhado popularidade com suas novas utilidades, e suas características costumam ser resumidas pelo 5 Vs: *volume*, *velocidade*, *variedade*, *veracidade* e *valor*. *Volume* em razão da quantidade massiva de dados gerados e processados; *velocidade*, pela rapidez com que tais dados são gerados e processados, em diferentes sistemas; *variedade*, pois há uma ampla diversidade de dados, sejam eles estruturados, semiestruturados ou não estruturados; *veracidade* no sentido da precisão e confiabilidade dos dados, que devem ser consistentes e corretos e gerados por um sistema confiável; e *valor*, relativo aos diferentes tipos de benefícios que podem ser auferidos a partir da utilização desses dados, incluindo ganhos econômicos (Younas, 2019).

Os dados podem ser compreendidos com uma das principais matérias-primas da economia na atualidade, que se pode chamar de *economia de dados*, e é a base de plataformas digitais e de empresas de tecnologia.

A internet é financiada pela coleta, análise e comércio de dados. Para tanto, as grandes corporações do Vale do Silício ofertam plataformas, redes sociais e aplicativos dos mais diversos tipos, que, por seu turno, oferecem serviços e “comodidades” em troca dos dados pessoais dos usuários. Tais dados, em seguida, são comercializados com empresas que os (re)utilizam para realizarem

⁵ Utilizamos aqui a nomenclatura adotada pela Lei Geral da Proteção de Dados Pessoais (LGPD) - Lei nº 13.709, de 14 de agosto de 2018, que assim o define: Art. 5º, VI - controlador: pessoa natural ou jurídica, de direito público ou privado, a quem competem as decisões referentes ao tratamento de dados pessoais.

⁶ Também com base na LGPD, titular é pessoa natural a quem se referem os dados pessoais que são objeto de tratamento (Art. 5º, V).

propagandas e marketing personalizados, em consonância com os perfis pessoais de cada usuário dos serviços digitais (Oliveira; Morais, 2021, p. 33).

Bem por isso, Morozov (2018) destaca algumas características e consequências dessa relativamente nova e lucrativa atividade base do que denominou de **capitalismo dadocêntrico**, uma visão de capitalismo que é centrada em uma lógica de constante maximização de lucro, muitas vezes associada à ideia de que tudo pode ser transformado em um ativo financeiro, até mesmo nossas atividades pessoais ou sociais que antes eram realizadas fora do ambiente de trabalho e distante das ambições do mercado⁷.

No mesmo sentido, Sant’Ana (2016) destaca que, no cenário do *Big Data*, a alta disponibilidade de dados oferece novas possibilidades e consequências, incluindo a necessidade de se criar soluções que conectem melhor os usuários às suas necessidades informacionais, o que é crucial para minimizar problemas como a **assimetria informacional**⁸, que pode ser prejudicial tanto para os indivíduos quanto para a sociedade, equilibrando assim os benefícios do acesso massivo a dados.

Com essa nova realidade trazida pelo *capitalismo dadocêntrico*, as diversas possibilidades relativas ao tratamento de dados, desde sua coleta, passando pelo armazenamento e processamento, ou seja, por todo o ciclo de vida dos dados, disseminaram-se para praticamente todas as áreas de forma atomizada: não é raro o recurso aos dados para tomadas de decisões no âmbito financeiro, comercial, da saúde, de políticas públicas, dentre tantos outros.

Bem por isso, no contexto de **protagonismo dos dados**, assim como diante da sua constante e veloz evolução, a Ciência de Dados aparece em posição de destaque enquanto área do conhecimento, de natureza interdisciplinar, destinada a efetuar o tratamento dos dados com o objetivo de auxiliar na tomada de decisões e solucionar problemas. A Ciência de Dados, portanto, pode ser compreendida como um “[...] conceito conexo à camada dos métodos, na qual os softwares são empregados para transformar

⁷ Nas palavras do autor: “O modelo de capitalismo “dadocêntrico” adotado pelo Vale do Silício busca converter todos os aspectos da existência cotidiana em ativo rentável: tudo aquilo que costumava ser o nosso refúgio contra os caprichos do trabalho e as ansiedades do mercado” (Morozov, 2018, p. 33).

⁸ “A expressão *information asymmetry* foi originalmente adotada pelo economista Akerlof (1970), quando da investigação das relações comerciais em que o vendedor de um bem tem o domínio das informações acerca deste, enquanto o comprador as desconhece” (Marques; Pinheiro, 2011, p. 72).

dados em informação, resultando no apoio à tomada de decisão” (Rautenberg; Carmo, 2019, p. 57) ou, de acordo com Provost e Fawcett (2013), representa um conjunto de princípios fundamentais que apoiam e orientam a extração de informação e conhecimento a partir de dados.

Até mesmo em razão da possibilidade de sua aplicação interdisciplinar, a Ciência de Dados, sobretudo quando conjugada com o uso da Inteligência Artificial, tem sido amplamente utilizada pelas mais diversas áreas, como por exemplo na oferta de produtos e de serviços de natureza financeira, com base na projeção de padrões relacionados a um público específico ou mesmo em relação às movimentações do mercado; na área de marketing, cuja análise dos dados relativos ao negócio permite alcançar uma maior efetividade na sua divulgação; e na área da saúde, já que a análise dos dados pode contribuir para o diagnóstico e para a prevenção de doenças, dentre outras tantas possibilidades. E, no âmbito do Direito, não tem ocorrido de maneira diferente.

3 CIÊNCIA DE DADOS E INTELIGÊNCIA ARTIFICIAL NO PODER JUDICIÁRIO BRASILEIRO

No contexto atual, em que os dados e sua exploração para os mais diversos fins fazem parte do cotidiano das pessoas, corporações e governos, inúmeros setores passaram a utilizá-los para auxiliá-los nas tomadas de decisões, seja mediante minuciosa análise e extração de informações a partir de gigantescos bancos de dados – *Big Data* – como também com a construção de modelos preditivos alimentados por Inteligência Artificial, com consideráveis níveis de precisão.

Diante dessa realidade, o Direito e as instituições jurídicas cederam à utilização de tais avanços tecnológicos a seu favor. As possibilidades são amplas: o uso da Ciência de Dados aliada à Inteligência Artificial pode contribuir com os profissionais do Direito (juízes, promotores, advogados etc.), por exemplo, com a finalidade de apoiar decisões (como na análise de precedentes e realização de previsões sobre decisões judiciais), auxiliar na apreciação de grandes bases de dados, facilitar a automação de processos e procedimentos jurídicos, dentre outras tantas utilidades.

O presente trabalho tem como objetivo analisar a aplicabilidade da Ciência de Dados aliada à Inteligência Artificial em um âmbito jurídico específico, qual seja, no Poder

Judiciário brasileiro. Para tal propósito, serão apresentados alguns dados obtidos no último mapeamento de projetos de Inteligência Artificial nos órgãos judiciários nacionais, mediante análise da “Pesquisa Uso de Inteligência Artificial (IA) no Poder Judiciário – 2023”, realizada e publicada pelo Conselho Nacional de Justiça (CNJ) no âmbito de 91 Tribunais e 3 Conselhos brasileiros.

De início, importa considerar que a pesquisa em exame teve por propósito:

[...] obter informações sobre tópicos de interesse diante das soluções que estão sendo criadas pelos órgãos judiciários, como o estado atual das iniciativas de IA, a identificação dos tipos de projetos de IA, as equipes que estão envolvidas nos projetos, os modelos de IA que estão sendo empregados e as implicações éticas e legais desses projetos (CNJ, 2024, p. 16).

Pelo estudo ora analisado, foi possível constatar o aumento do uso de Inteligência Artificial pelos órgãos do Poder Judiciário de nosso país, seja por meio de projetos já implementados ou em andamento, além de redução de projetos não iniciados, quando em comparação com o período anterior (comparou-se com a pesquisa relativa ao ano anterior, de 2022, também realizada pelo CNJ). Do total das instituições estudadas e que responderam à pesquisa, foi verificado a existência de 140 projetos de IA⁹, sendo que 45% deles já estão em desenvolvimento, logo, com alguma aplicação prática:

Dos tribunais respondentes, 62 (66% do total) informaram ter projetos de IA em desenvolvimento: 63 projetos (45% do total) estão em produção, ou seja, já há uma aplicação prática da IA no cotidiano judiciário; 17 (12,1% do total) encontram-se em estágio inicial; 46 (32,9% do total) estão em andamento; e 11 (7,9% do total) foram finalizados. Por fim, 3 projetos citados nas respostas (2,1% do total) foram declarados como não iniciados (CNJ, 2024, p. 27).

Nota-se, portanto, que o uso da Inteligência Artificial nos Tribunais brasileiros é uma realidade, e que tem aumentado a cada ano, inclusive. Os motivos elencados para o uso da IA são os mais variados, e pela pesquisa do CNJ foi possível notar que a gestão da alta demanda de processos é uma questão central na implementação de soluções de IA, especialmente nas Justiças Estadual e Federal como também nos Tribunais Superiores.

⁹ “Os projetos de IA nos tribunais cobrem uma ampla variedade de atividades, desde a automação de processos administrativos até análises avançadas de dados, visando otimizar a eficiência judicial e melhorar o acesso à Justiça. Além disso, incorporam uma série de automações para enfrentar os desafios impostos pelo grande volume de casos e documentos” (CNJ, 2024, p. 57).

No mesmo sentido, a distribuição dos projetos envolvendo uso de Inteligência Artificial também apresentou variações importantes, de modo que houve um

[...] envolvimento mais acentuado da Justiça Estadual, com 68 projetos (48,6% do total), seguida pela Eleitoral, com 23 (16,4% do total), e do Trabalho, com 20 (14,3% do total) (CNJ, 2024, p. 27).

Por sua vez, “O Judiciário federal e os tribunais superiores também participaram, com 14 (10% do total) e 13 projetos (9,3% do total), respectivamente” (CNJ, 2024, p. 28) e, “no âmbito dos Conselhos (Conselho da Justiça Federal – CJF e Conselho Nacional de Justiça - CNJ), verificou-se haver 2 projetos, que representam 1,4% do total” (CNJ, 2024, p. 27).

Muito embora a quantidade de projetos na Justiça Estadual seja numericamente maior, quando comparado com o número de Tribunais que compõem a estrutura do Poder Judiciário em seus diferentes ramos e especialidades, o percentual revela que, embora os Tribunais Estaduais contem com o maior número de projetos, não significa necessariamente que estejam mais empenhados no uso da IA. Os números demonstram que, proporcionalmente, os Tribunais Superiores apresentam maior média de projetos de IA por Tribunal (embora tais índices estejam próximos aos órgãos federais e estaduais), seguidos pelos Tribunais Federais, Estaduais, Eleitorais, do Trabalho e pelos Conselhos, nesta sequência.

Os tribunais superiores, com 13 projetos distribuídos por 5 tribunais, apresentam uma média de 2,6 projetos por tribunal, sugerindo uma alta densidade de projetos de IA em comparação com outros ramos. Da mesma forma, os tribunais federais, com 14 projetos em 6 tribunais, têm uma média de 2,3 projetos por tribunal, indicando também uma forte adesão à IA. Enquanto isso, os 27 tribunais estaduais, apesar de contarem com o maior número absoluto de projetos (68), têm uma média de 2,5 projetos por tribunal, que é comparável à dos tribunais superiores e federais. Os 27 tribunais eleitorais e 24 do Trabalho, com 23 e 20 projetos, respectivamente, apresentam médias menores: 0,85 e 0,83 projeto por tribunal. Os 3 conselhos, com 2 projetos, têm uma média de 0,67 projeto por conselho (CNJ, 2024, p. 47).

Observa-se também a existência de Tribunais e um Conselho que relataram não possuir projetos de IA. A contagem por Tribunais revela os seguintes números de órgãos carentes de projetos dessa natureza: no ramo Eleitoral, 11 Tribunais, representando 40,7% dos Tribunais Eleitorais; no ramo Trabalhista, 9 Tribunais, ou seja, 37,5% dos

Tribunais da Justiça do Trabalho; no ramo Estadual, 6 Tribunais, correspondendo a 22,2% dos Tribunais Estaduais; no ramo Federal, 2 Tribunais, que correspondem a 33,3% dos Tribunais Federais; no ramo Militar, 2 Tribunais, ou 66,7% dos Tribunais Militares; no ramo Superior, 1 Tribunal, equivalente a 25% dos Tribunais Superiores; e, por fim, 1 órgão dos Conselhos, representando 33,3% dos Conselhos (CNJ, 2024).

Assim, com exceção da Justiça Militar, no contexto de todos os demais Tribunais da Justiça Comum e Especializada, bem como dos Tribunais Superiores e Conselhos, em mais de 50% da sua totalidade foi iniciado algum projeto envolvendo a utilização de Inteligência Artificial.

Outro ponto a ser observado diz respeito aos benefícios elencados pelos entrevistados a respeito do uso da IA no Poder Judiciário, sendo eficiência e agilidade no processamento de documentos o mais apresentado nas respostas:

Os benefícios mais citados da implementação de IA incluem maior eficiência e agilidade ao processar documentos, com 74 citações, em 52,8% das respostas; otimização de recursos, com 68 menções, em 48,6% das respostas; automatização de tarefas repetitivas, com 63 menções, em 45% das respostas; e redução do tempo de tramitação dos processos, com 52 citações, em 37,1% das respostas (CNJ, 2024, p. 34).

Há mais um ponto que merece destaque, dentre tantos fatores importantes expostos pela pesquisa em discussão, qual seja, o estabelecimento de parcerias para os desenvolvimentos dos projetos de IA. Constatou-se que 65 projetos, correspondentes a 46,4% do total, foram realizados sem o apoio de outras entidades; 39 projetos, ou 27,9% do total, foram realizados em parceria com outras entidades jurídicas, como tribunais, órgãos ou entidades jurídicas; 23 projetos, representando 16,4% do total, foram desenvolvidos em colaboração com universidades; 13 projetos, equivalentes a 9,3% do total, foram feitos em parceria com a iniciativa privada; e 3 projetos, ou 2,1% do total, contaram com o apoio de outras entidades públicas, não necessariamente da área jurídica (CNJ, 2024). Esse cenário demonstra uma cooperação entre os órgãos do Poder Judiciário entre si, mas também com apoio de setores externos, como a iniciativa privada e outras entidades – jurídicas ou não, destacando-se também o recurso às universidades.

Por fim, cabem algumas menções acerca dos riscos técnicos e preocupações éticas relativos ao uso da IA, e que podem interferir na realização ou na continuidade de

projetos de IA pelos entes judiciários. No âmbito técnico, foram destacados pontos importantes sobre o risco de inconsistência de dados associado ao uso limitado de uma plataforma eletrônica centralizada, razão pela qual a maioria dos tribunais optou por utilizar uma infraestrutura própria, em vez de contratar serviços de nuvem, para garantir a privacidade dos dados sensíveis dos documentos jurídicos. Já em relação às questões éticas, elas podem ser consideradas junto com outros desafios gerais enfrentados para a não implementação de projetos, já que essa preocupação foi manifestada por todos os tribunais.

A pesquisa é extensa e bastante completa, com diversas constatações e valorosas conclusões que mapeiam de maneira significativa o atual estágio do uso da Inteligência Artificial pelos órgãos do Poder Judiciário brasileiro. No presente estudo, foram abordadas apenas algumas das conclusões extraídas da pesquisa do Conselho Nacional de Justiça, consideradas pertinentes ao que se propõe. Contudo, é crucial salientar que a pesquisa como um todo contém uma vasta gama de dados e informações relevantes, impossíveis de serem esgotados nas limitações deste artigo.

É certo que a Inteligência Artificial consiste em uma tecnologia em constante evolução, que pretende perdurar e alcançar os mais diversos setores da sociedade. Embora já exista há algumas décadas, não há como negar que nos últimos anos seu aprimoramento foi surpreendente, cujas aplicações são aptas a proporcionar um nível significativo de facilidades e benefícios para os indivíduos e empresas. Da mesma forma, pode ocasionar problemas e danos diretamente proporcionais aos ganhos. Seu uso ético e responsável, portanto, deve ser assunto constantemente debatido.

No contexto de uma infinidade de dados – *Big Data* – que são coletados e processados com finalidades diversas, especialmente econômicas, diferentes setores já utilizam a IA com o escopo precípua de aprimorar seus processos e otimizar o emprego de recursos financeiros. Bem por isso, no âmbito jurídico ela já vem sendo utilizada por praticamente todos os sujeitos nele envolvidos, incluindo o Poder Judiciário.

Por tal motivo, procedeu-se a uma análise da *Pesquisa Uso de Inteligência Artificial (IA) no Poder Judiciário – 2023*, de iniciativa do Conselho Nacional de Justiça e realizada no âmbito de 91 Tribunais e 3 Conselhos brasileiros, com o objetivo de extrair

uma visão atual sobre a utilização de tal tecnologia no âmbito dos órgãos judiciários brasileiros.

A partir de uma investigação sobre o uso da IA com base nas informações encontradas na mencionada pesquisa, embora limitada pelas circunstâncias deste estudo, pôde-se constatar que essa prática já é uma realidade nos Tribunais brasileiros. Dentre as importantes constatações trazidas pela pesquisa, destacam-se as seguintes, iniciando pela amplitude de sua utilização:

- a) 62 tribunais (66% do total) informaram ter projetos de IA em desenvolvimento;
- b) 63 projetos (45% do total) estão em produção, com aplicação prática da IA no cotidiano judiciário;
- c) 17 projetos (12,1% do total) estão em estágio inicial;
- d) 46 projetos (32,9% do total) estão em andamento;
- e) 11 projetos (7,9% do total) foram finalizados;
- f) 3 projetos (2,1% do total) foram declarados como não iniciados.

Em relação à distribuição nos diversos ramos do Poder Judiciário, verificou-se que:

- a) Nos Tribunais Superiores, há 13 projetos distribuídos por 5 tribunais, que representa uma média de 2,6 projetos por tribunal, logo, uma alta densidade de projetos de IA;
- b) nos Tribunais Federais, há 14 projetos em 6 tribunais, sendo uma média de 2,3 projetos por tribunal, ou seja, há forte adesão à IA;
- c) nos Tribunais Estaduais, são 68 projetos em 27 tribunais, com média de 2,5 projetos por tribunal, comparável aos tribunais superiores e federais;
- d) nos Tribunais Eleitorais, há 23 projetos em 27 tribunais, com média de 0,85 projeto por tribunal;
- e) nos Tribunais do Trabalho, são 20 projetos em 24 tribunais, e uma média de 0,83 projeto por tribunal;

- f) por fim, nos Conselhos, há 2 projetos em 3 conselhos, com média de 0,67 projeto por conselho – representando as menores médias juntamente com os Tribunais Eleitorais e do Trabalho.

No estudo também se verificou que existem órgãos que não possuem projetos de IA, sendo os com mais abstenções os componentes das Justiças Especializadas Militar e Eleitoral, a saber:

- a) 11 Tribunais Eleitorais sem projetos de IA (40,7% do total);
- b) 09 Tribunais Trabalhistas sem projetos de IA (37,5% do total);
- c) 06 Tribunais Estaduais sem projetos de IA (22,2% do total);
- d) 02 Tribunais Federais sem projetos de IA (33,3% do total);
- e) 02 Tribunais Militares sem projetos de IA (66,7% do total);
- f) 01 Tribunal Superior sem projetos de IA (25% do total);
- g) 01 Conselhos sem projetos de IA (33,3% do total).

Dentre os benefícios mais citados à implementação de IA pelos Tribunais, destacam-se a eficiência e agilidade no processamento de documentos como a mais citada:

- a) Maior eficiência e agilidade ao processar documentos (74 citações, 52,8% das respostas);
- b) otimização de recursos (68 menções, 48,6% das respostas);
- c) automatização de tarefas repetitivas (63 menções, 45% das respostas);
- d) redução do tempo de tramitação dos processos (52 citações, 37,1% das respostas).

A cooperação dos órgãos judiciários entre si e com outros setores, jurídicos e não jurídicos, também mereceu destaque, já que se pôde observar uma cooperação significativa entre órgãos do Poder Judiciário e setores externos, incluindo iniciativa privada e universidades:

- a) 65 projetos realizados sem apoio de outras entidades (46,4% do total);
- b) 39 projetos realizados em parceria com outras entidades jurídicas (27,9% do total);

- c) 23 projetos desenvolvidos em colaboração com universidades (16,4% do total);
- d) 13 projetos feitos em parceria com a iniciativa privada (9,3% do total);
- e) 03 projetos com apoio de outras entidades públicas, não necessariamente jurídicas (2,1% do total). (CNJ, 2024).

Por último, há menções acerca dos riscos técnicos e preocupações éticas relativos ao uso da IA, aptos a interferir no início e desenvolvimentos de projetos dessa natureza.

4 CONSIDERAÇÕES FINAIS

O recurso à Inteligência Artificial já é uma realidade nos Tribunais brasileiros, com destaque à amplitude de sua utilização (66% dos Tribunais que responderam à pesquisa informou ter projetos de IA em desenvolvimento), com uma maior média de projetos nos Tribunais Superiores. Quanto aos benefícios do uso da IA, constatou-se que a busca por maior eficiência e agilidade ao processar documentos foi a resposta mais apresentada. Também se verificou que há uma cooperação significativa entre órgãos do Poder Judiciário e setores externos, incluindo iniciativa privada e universidades, para a realização de tais projetos. Por fim, foram mencionadas preocupações com os riscos técnicos e receios de natureza ética relacionados ao uso da IA, que podem impactar no início e o desenvolvimento dos projetos dessa categoria.

Conclui-se, portanto, que atualmente há uma alta adesão ao uso de Inteligência Artificial no Poder Judiciário brasileiro, com tendência de ampliação, embora tais tecnologias ainda estejam em desenvolvimento e existam preocupações quanto à sua implementação.

REFERÊNCIAS

ALVES, F. G.; ALVES, V. R. F.; MENEZES, R. O. A.; MOREIRA, T. O. (org.). **Direito aplicado: inovação e sociedade**. Natal: Polimatia, 2023. v. 2.

ASSIS, Cristina Ferreira; MONTEIRO, Rhadson. Metodologias qualitativas e quadros de referência para a pesquisa em Ciências Humanas e Sociais Aplicadas. **Revista Jures**, [S.

/], v. 16, n. 29, p. 1-28, 2023. Disponível em:
<https://estacio.periodicoscientificos.com.br/index.php/juresvitoria/article/view/1993>.
Acesso em: 4 maio 2025.

BODEN, M. A. **Inteligência artificial**: uma brevíssima introdução. São Paulo: Editora Unesp, 2020.

BRASIL. Lei nº 13.709, de 14 de agosto de 2018. Lei Geral da Proteção da Dados Pessoais (LGPD). Dispõe sobre a proteção de dados pessoais e altera a Lei nº 12.965, de 23 de abril de 2014 (Marco Civil da Internet). **Diário Oficial da União**: seção 1, Brasília, DF, n. 157, p. 59, 15 ago. 2018. Disponível em:
https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709compilado.htm.
Acesso em: 1 ago. 2024.

CONSELHO NACIONAL DE JUSTIÇA (CNJ). **Pesquisa uso de inteligência artificial (IA) no poder judiciário 2023**. Brasília: CNJ, 2024. Disponível em: <https://www.cnj.jus.br/wp-content/uploads/2024/06/pesquisa-uso-da-inteligencia-artificial-ia-no-poder-judiciario-2023.pdf>. Acesso em: 4 maio 2025.

KAUFMAN, Dora. **A inteligência artificial irá suplantará a inteligência humana?**. Barueri: Estação das Letras e Cores, 2019.

MARQUES, Rodrigo Moreno; PINHEIRO, Marta Macedo Kerr. Política de informação nacional e assimetria de informação no setor de telecomunicações brasileiro. **Perspectivas em Ciência da Informação**, [S. l.], v. 16, n. 1, p. 65-91, jan./mar. 2011. Disponível em: <https://periodicos.ufmg.br/index.php/pci/article/view/22714/18298>. Acesso em: 4 maio 2025.

MOROZOV, Evgeny. **Big Tech**: ascensão dos dados e a morte da política. São Paulo: Ubu Editora, 2019.

PROVOST, Foster; FAWCETT, Tom. Data science and its relationship to big data and data-driven decision making. **Big data**, [S. l.], v. 1, n. 1, p. 51-59, 2013. Disponível em: <https://fosterprovost.com/wp-content/uploads/2019/07/Data-science-and-its-relationship-to-big-data-and-data-driven-decision-making.pdf>. Acesso em: 4 maio 2025.

RAUTENBERG, S.; CARMO, P. R. V. Big data e ciência de dados: complementariedade conceitual no processo de tomada de decisão. **Brazilian Journal of Information Science: research trends**, Marília, v. 13, n. 1, p. 56-67, 2019. Disponível em: <https://revistas.marilia.unesp.br/index.php/bjis/article/view/8315>. Acesso em: 20 nov. 2020.

SANT'ANA, Ricardo César Gonçalves. Ciclo de vida dos dados: uma perspectiva a partir da Ciência da Informação. **Inf. Inf.**, Londrina, v. 21, n. 2, p. 116-142, maio/ago. 2016. Disponível em: <https://ojs.uel.br/revistas/uel/index.php/informacao/article/view/27940/20124>. Acesso em: 4 maio 2025.

SETZER, Valdemar W. **Dado, informação, conhecimento e competência**. São Paulo: USP, 2015. Disponível em: <https://www.ime.usp.br/~vwsetzer/dado-info.html>. Acesso em: 13 ago. 2024.

YOUNAS, Muhammad. Research challenges of Big Data. **Service Oriented Computing and Applications**, [S. l.], v. 13, n. 2, p. 105-107, 2019. Disponível em: <https://link.springer.com/article/10.1007/s11761-019-00265-x>. Acesso em: 4 maio 2025.



[4.0 Internacional](https://creativecommons.org/licenses/by/4.0/)

Esta licença permite compartilhamento, remixe, adaptação e criação a partir do trabalho, mesmo para fins comerciais, desde que sejam atribuídos créditos ao(s) autor(es). Conteúdos elaborados por terceiros, citados e referenciados nesta obra não são cobertos pela licença.

Como citar este trabalho:

PEREIRA, Fabrício Fracaroli; SABBAG, Deise Maria Antonio. *Ciência de Dados e Inteligência Artificial no Poder Judiciário brasileiro*. In: WORKSHOP DE INFORMAÇÃO DADOS E TECNOLOGIA, 8., 2025, Marília, SP. **Anais [...]**. Marília, SP: Universidade de Marília, 2025. DOI: <https://doi.org/10.22477/viii.widat.270>.



WIDaT 2025

VIII WORKSHOP DE INFORMAÇÃO,
DADOS E TECNOLOGIA

MARÍLIA • SP

Mapeamento de dados como arcabouço para a Arquitetura de Dados da Avaliação da Pós-Graduação Brasileira no contexto da CAPES

Data mapping as a framework for the Data Architecture of the Brazilian Postgraduate Evaluation in the context of CAPES

Emanuelle Torino¹



<http://lattes.cnpq.br/5042296869081637>



<https://orcid.org/0000-0002-3791-9884>

Caio Saraiva Coneglian²



<http://lattes.cnpq.br/3954065076810604>



<https://orcid.org/0000-0002-6126-9113>

Mariana Baptista Brandt³



<http://lattes.cnpq.br/3761037263199030>



<https://orcid.org/0000-0001-8119-7527>

Talita Moreira de Oliveira⁴



<http://lattes.cnpq.br/2374823276591654>



<https://orcid.org/0000-0001-5740-4207>

Silvana Aparecida Borsetti Gregorio Vidotti⁵



<http://lattes.cnpq.br/7390573927636069>



<https://orcid.org/0000-0002-4216-0374>

¹ Doutora em Ciência da Informação, Universidade Estadual Paulista Júlio de Mesquita Filho (Unesp), Marília, São Paulo, Brasil. Bibliotecária da Universidade Tecnológica Federal do Paraná (UTFPR), Curitiba, Paraná, Brasil. Coordenadora da Coordenação de Serviços Bibliográficos (COBIB), Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), Brasília, Distrito Federal, Brasil. emanuelle@utfpr.edu.br.

² Doutor em Ciência da Informação, Universidade Estadual Paulista Júlio de Mesquita Filho (Unesp), Marília, São Paulo, Brasil. Professor e coordenador dos cursos de graduação de Análise e Desenvolvimento de Sistemas e Ciência da Computação da Universidade de Marília (Unimar), Marília, São Paulo, Brasil. Professor do Programa de Pós-graduação em Ciência da Informação, Universidade Estadual Paulista Júlio de Mesquita Filho (Unesp), Marília, São Paulo, Brasil. caio.coneglian@gmail.com.

³ Doutora em Ciência da Informação, Universidade Estadual Paulista Júlio de Mesquita Filho (Unesp), Marília, São Paulo, Brasil. Analista Legislativo - Documentação e informação legislativa na Câmara dos Deputados, Brasília, DF, Brasil. mariana.brandt@unesp.br.

⁴ Doutora em Educação em Ciências, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Rio Grande do Sul, Brasil. Analista em Ciência e Tecnologia na Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brasília, Distrito Federal, Brasil. talita.oliveira@capes.gov.br.

⁵ Doutora em Educação, Universidade Estadual Paulista Júlio de Mesquita Filho (Unesp), Marília, São Paulo, Brasil. Professora do Departamento e do Programa de Pós-graduação em Ciência da Informação, Universidade Estadual Paulista Júlio de Mesquita Filho (Unesp), Marília, São Paulo, Brasil. Coordenadora de Tecnologias Aplicadas (COTEA), Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), Brasília, Distrito Federal, Brasil. silvana.vidotti@unesp.br.

Resumo

O Ecossistema de Avaliação da Pós-Graduação Brasileira busca aprimorar a disponibilização de dados para processos como a avaliação dos Programas de Pós-Graduação *stricto sensu*, o acompanhamento institucional e a transparência para a sociedade. No entanto, a diversidade estrutural dos sistemas de informação utilizados tem dificultado a padronização e a interoperabilidade dos dados, impactando sua integração e a tomada de decisão. Este estudo objetiva apresentar o mapeamento dos dados da avaliação da pós-graduação *stricto sensu* brasileira no contexto da CAPES, fornecendo subsídios para o desenvolvimento de uma Arquitetura de Dados como alicerce à integração de dados. Trata-se de uma pesquisa de natureza empírica utilizando-se de fonte documental, que investiga um fenômeno com escassez de literatura disponível em âmbito nacional e internacional, evidenciando sua relevância. O estudo fundamenta-se em análise exploratória da estrutura de dados alicerçada em referenciais teóricos e práticos das áreas de Ciência da Informação, Sistemas de Informação e Ciência da Computação. A metodologia adotada consistiu em uma análise detalhada da estrutura de dados utilizada na avaliação da pós-graduação, com ênfase na integração e interoperabilidade dos dados. Como resultado, obteve-se o mapeamento da estrutura de dados da avaliação da pós-graduação *stricto sensu* brasileira no contexto da CAPES, que servirá como base para o desenvolvimento de uma Arquitetura de Dados para esse processo.

Palavras-chave: Arquitetura de Dados; mapeamento de dados; integração de dados; avaliação da pós-graduação.

Abstract

The Brazilian Graduate Evaluation Ecosystem aims to improve data availability for processes such as the evaluation of *stricto sensu* Graduate Programs, institutional monitoring, and transparency for society. However, the structural diversity of the information systems used makes it difficult for data standardization and interoperability, impacting their integration and decision-making processes. This study's objective is to present the mapping of evaluation data of Brazilian *stricto sensu* graduate programs in the context of CAPES, providing support for the development of a Data Architecture to serve as the foundation for data integration. It is an empirical research using documentary sources, that investigates a phenomenon with limited available literature both nationally and internationally, highlighting its relevance. The study is based on an exploratory analysis of the data structure, grounded in theoretical and practical frameworks from the fields of Information Science, Information Systems, and Computer Science. The adopted methodology consisted of a detailed analysis of the data structure used in graduate evaluation, with an emphasis on data integration and interoperability. As a result, the study achieved the mapping of the data structure for the evaluation of Brazilian *stricto sensu* graduate programs in the CAPES context, which will serve as the basis for the development of a Data Architecture for this process.

Keywords: Data Architecture; data mapping; data integration; postgraduate evaluation.

1 INTRODUÇÃO

O ecossistema de avaliação da pós-graduação brasileira tem se desenvolvido buscando favorecer as formas de disponibilização dos dados para diferentes processos, dentre os quais destacam-se: a avaliação dos programas de pós-graduação *stricto sensu*, o acompanhamento do desenvolvimento destes programas pelas próprias instituições e a abertura de dados para a sociedade.

Diferentes sistemas de informação foram utilizados ao longo dos anos, com vistas a favorecer o processamento dos dados pelas instituições envolvidas na avaliação

da pós-graduação *stricto sensu* brasileira, seja por instituições de ensino, agências de fomento e pela própria Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). Desta forma, a heterogeneidade estrutural desses sistemas apresenta desafios à padronização e integração de dados, o que dificulta a interoperabilidade entre as plataformas, além de tornar mais complexo o processo de tomada de decisão baseado em dados.

O estudo de Nóbrega, Torino, Vidotti, Halicki e Oliveira (2023, p. 2) observou que “No ecossistema brasileiro de Educação, Ciência, Tecnologia e Inovação (ECTI), a coleta de dados armazenados em diferentes sistemas de informação é complexa em função das estruturas e formas de armazenamento dos dados.” Além disso, o estudo possibilitou o desenvolvimento, pela CAPES, de Ferramentas de Apoio à Integração de Sistemas (FAIS) como solução para a integração de dados da avaliação da pós-graduação *stricto sensu* brasileira.

No entanto, tais ferramentas dependem da compreensão das estruturas de dados entre as instituições que oferecem programas de pós-graduação *stricto sensu* e a própria CAPES, que atuam neste processo, respectivamente como fornecedora e consumidora de dados.

Desta forma, este trabalho apresenta uma parte de um projeto de pesquisa que busca desenvolver a Arquitetura de Dados da Avaliação da Pós-Graduação Brasileira no contexto da CAPES. A proposição de tal Arquitetura apoiará a compreensão das estruturas de dados para otimizar os processos de integração.

Para tanto, é necessário compreender a estrutura atual dos dados e avaliá-la para que a Arquitetura de Dados a ser desenvolvida possa refletir a necessidade da CAPES no que tange à avaliação da pós-graduação *stricto sensu* brasileira.

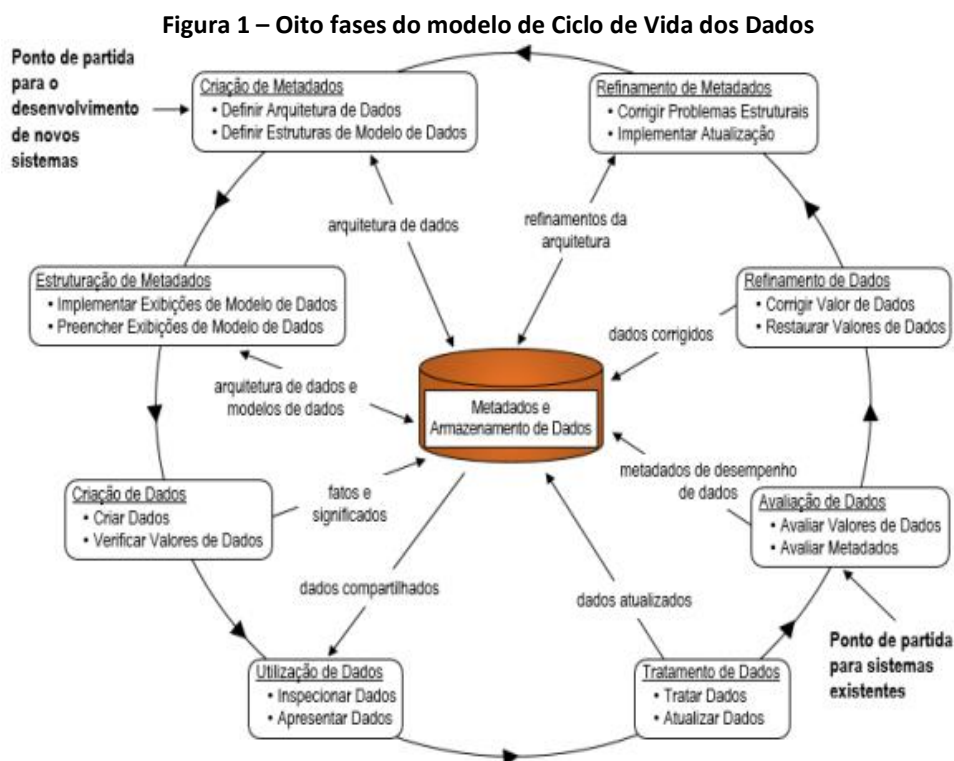
Tem-se assim, como objetivo deste estudo apresentar o mapeamento dos dados da avaliação da pós-graduação *stricto sensu* brasileira no contexto da CAPES, visando embasar a Arquitetura de Dados a ser desenvolvida como alicerce à integração de dados.

1.1 Arquitetura de Dados

Neste estudo, o arcabouço conceitual utilizado provém da Arquitetura de Dados no contexto da Ciência da Informação, conceituada por Torino (2023, p. 192) como:

A Arquitetura de Dados atua no projeto da infraestrutura de dados e do ciclo de vida dos dados, de forma holística e coesa, considerando que os dados necessitam de padrões de metadados, estruturas sintática e semântica e atendam a princípios éticos, legais e técnicos para que sejam consumidos e fornecidos para diferentes finalidades utilizando técnicas e tecnologias emergentes.

Destaca-se que a Arquitetura de Dados pode ser desenvolvida quando do planejamento de um sistema de informação ou considerando a estrutura de dados existente em um sistema já implantado. No âmbito da Arquitetura de Dados, destaca-se o ciclo de vida dos dados que, de acordo com o modelo proposto por Yoon, Aiken e Guimarães (2005), é constituído por oito fases (Figura 1) que podem ser iniciadas pelo "Ponto de partida para o desenvolvimento de novos sistemas" ou pelo "Ponto de partida para sistemas existentes".



Fonte: Yoon, Aiken e Guimarães (2005, p. 147) traduzido por Torino (2023, p. 198).

Neste estudo, considerando a existência de uma infraestrutura de sistemas de informação e uma estrutura de dados relativos ao processo de avaliação da pós-graduação *stricto sensu* brasileira no contexto da CAPES, adotou-se o "Ponto de partida para sistemas existentes". Neste cenário, o processo foi iniciado pela avaliação dos dados e, considerando a Arquitetura de Dados, a avaliação se aplicou à estrutura de dados composta por metadados, sobretudo aos atributos.

Os atributos são relevantes, uma vez que a Arquitetura de Dados projeta os dados para que sejam facilmente consumidos por sistemas de informação, desde que atendam aos requisitos estabelecidos (Torino, 2023). Assim, dados normalizados aumentam a possibilidade de compartilhamento, em especial quando orientados pelo uso e não pela aplicação (Tupper, 2011).

Essa abordagem reforça que a necessidade de integração de sistemas tem sido uma problemática que culminou na alteração do desenvolvimento de sistemas, cuja abordagem de sistemas isolados passa a dar lugar ao suporte multifuncional, que requer tratamento de dados para assegurar sua qualidade, integração e interoperabilidade (Yoon; Aiken; Guimarães, 2005). Destaca-se que, para que as integrações sejam satisfatórias, é necessário que a compreensão dos dados seja clara tanto para humanos quanto para aplicações computacionais.

Neste sentido, Torino (2023) afirma que a Arquitetura de Dados pode contribuir para a melhoria da sintaxe e da semântica dos dados, necessárias para a compreensão do sentido dos dados, para otimizar a comunicação dos dados em fluxos de integração, interoperabilidade, processamento, uso e reuso.

Desta forma, considerando que o processo de avaliação da pós-graduação *stricto sensu* no contexto da CAPES já possui uma estrutura de dados e sistemas existentes, tornou-se necessário mapear a estrutura de dados disponível, visando avaliá-la para que os refinamentos e correções necessários possam ser considerados na proposição de uma Arquitetura de Dados.

2 METODOLOGIA E MAPEAMENTO DOS DADOS

O presente estudo caracteriza-se como uma pesquisa empírica, que investiga a estrutura de dados da avaliação da pós-graduação *stricto sensu* brasileira, com foco na CAPES.

A pesquisa adota uma abordagem exploratória, pautada na análise da estrutura de dados utilizada na avaliação da pós-graduação, alicerçada em referenciais teóricos e práticos das áreas de Ciência da Informação, Sistemas de Informação e Ciência da Computação, bem como de fontes documentais. A análise foi conduzida a partir da coleta, tratamento e consolidação dos dados da CAPES, com o objetivo de identificar padrões, redundâncias e oportunidades para aprimoramento da integração e interoperabilidade dos dados.

Para realizar o processo do mapeamento dos dados de avaliação da pós-graduação, identificou-se, no contexto da CAPES, documentos e bancos de dados disponíveis que são capazes de apoiar a compreensão dos atributos e dos metadados que são parte do cenário tratado no âmbito desta pesquisa.

Inicialmente, para a realização do mapeamento dos dados de avaliação da pós-graduação brasileira, buscou-se compreender o estado atual e os avanços que a CAPES realizou nos últimos anos no que tange à integração de dados com as Instituições de Ensino Superior.

Neste cenário, em 2023, a CAPES lançou o Programa de Governança Colaborativa de Informações da Pós-Graduação (GoPG) para promover a interoperabilidade entre sistemas acadêmicos, científicos, tecnológicos e de inovação, além de fornecer ferramentas de gestão da informação e estabelecer padrões e serviços com controle e proteção de dados (Brasil, 2024). Integrada ao GoPG está a Rede de Integração da Comunidade Acadêmico-Científica | Pós-Graduação (RICA|PG), que tem como missão a construção de conexões para compartilhamento e reuso de dados da pós-graduação, sendo responsável pela definição de padrões para a comunicação de dados no processo de avaliação da pós-graduação *stricto sensu* no Brasil.

Considerando a missão da RICA|PG e a necessidade de desenvolvimento de uma Arquitetura de Dados para a avaliação da pós-graduação *stricto sensu* no contexto da

CAPES, iniciou-se o mapeamento dessa estrutura de dados, visando embasar a Arquitetura de Dados a ser desenvolvida como alicerce à integração de dados.

A partir da compreensão deste cenário, iniciou-se o processo de mapeamento, cuja coleta dos dados ocorreu entre outubro de 2023 e janeiro de 2024, utilizando-se fontes documentais oficiais disponibilizadas pela CAPES. O mapeamento considerou as seguintes fontes de dados: Inventário de Dados da Plataforma Sucupira, Proposta de Entidades derivadas do "Relatório da Rede RICA|PG", Dados Abertos da CAPES, Dicionário de Dados da Plataforma Sucupira, descritas a seguir.

Inventário de Dados da Plataforma Sucupira: elaborado com base na interface de uso da Plataforma Sucupira, fornece uma visão detalhada e abrangente dos conjuntos de dados, metadados e variáveis utilizados na Plataforma Sucupira. Esse inventário contém informações específicas sobre os dados coletados e armazenados pela CAPES. Desta fonte foram extraídos 509 atributos e incluídos na análise 438 atributos, que fornecem uma base para compreender a estrutura dos dados e as informações que são gerenciadas pela Plataforma Sucupira.

Proposta de Entidades derivadas do "Relatório da Rede RICA|PG":, representam uma fonte importante, pois refletem o trabalho desenvolvido pelas instituições que compõem a RICA|PG com o intuito de definir entidades e variáveis relevantes para a avaliação da pós-graduação *stricto sensu* brasileira, sob a ótica das Instituições que possuem programas de pós-graduação. Desta fonte, foram extraídos 246 atributos e incluídos na análise 153 atributos, que representam as diferentes dimensões e aspectos definidos pelas instituições partícipes da RICA|PG.

Dados Abertos da CAPES: fornecem uma perspectiva dos dados da avaliação da pós-graduação *stricto sensu* brasileira que são de interesse da comunidade, uma vez que são exibidos publicamente no Portal Dados Abertos CAPES. Desta fonte, foram coletados 1.443 atributos, dos quais foram incluídos na análise 330 atributos únicos.

Dicionário de Dados da Plataforma Sucupira: consiste nos esquemas e atributos presentes no banco de dados da Plataforma Sucupira, de forma que representa uma fonte fundamental que proporciona uma visão detalhada e estruturada dos atributos e variáveis presentes no referido banco. Desta fonte foram extraídos 8.411 atributos que

oferecem uma compreensão abrangente da estrutura e do conteúdo do banco de dados da Plataforma Sucupira. Para a análise foram incluídos 854 atributos únicos, visando contribuir para uma visualização detalhada e precisa.

No total, foram mapeados mais de 10 mil atributos das quatro fontes citadas, que, após processos de normalização e agrupamento, resultaram em 1.775 atributos únicos representativos do processo de avaliação da pós-graduação *stricto sensu*, necessários para compreender a estrutura de dados utilizada para a avaliação da pós-graduação brasileira no contexto da CAPES.

Destaca-se que a totalidade dos atributos disponíveis em cada fonte de dados foi extraída, no entanto, foram realizadas análises prévias para identificar aqueles que são utilizados no processo investigado neste trabalho. Além disso, houve um tratamento para que os atributos que estão presentes em mais de uma entidade fossem agrupados, reduzindo o quantitativo, por serem considerados o mesmo atributo, por isso, designados atributos únicos.

Tendo em vista as especificidades de cada fonte de dados, realizou-se um processo de extração para que as estruturas pudessem ser compreendidas e registradas de forma padronizada. Para tanto, foi desenvolvido um "Instrumento de Análise de Dados da Avaliação da Pós-Graduação Brasileira", utilizado para identificar os atributos únicos relacionados ao processo em análise. Tal instrumento foi construído utilizando como ferramenta o Google Planilhas. Aponta-se ainda que as terminologias adotadas em cada uma das fontes de dados analisadas foram mantidas, visando elucidar questões semânticas que estão relacionadas ao processo de avaliação da pós-graduação pela CAPES.

3 RESULTADOS

O "Instrumento de Análise de Dados da Avaliação da Pós-Graduação Brasileira" foi desenvolvido para registrar os elementos presentes em cada uma das fontes de dados analisadas. Para tanto, foram criadas categorias correspondentes às entidades do Modelo Conceitual da Plataforma Sucupira, considerado elemento central para a compreensão da estrutura de dados da avaliação da pós-graduação *stricto sensu* brasileira, composto por

Programa, Área de concentração, Linha de pesquisa, Projeto, Produção intelectual, Tese, Docente, Discente, Participante externo, Disciplina e Curso. A partir delas, as entidades, conjuntos de informações, metadados, variáveis e outros elementos relevantes, provenientes das quatro fontes de dados foram categorizadas conforme o tema das categorias e registradas em um único documento em um processo de compatibilização, considerando a correspondência dos atributos entre as fontes.

A Figura 2 apresenta um recorte do instrumento, que demonstra como as 4 fontes foram alinhadas e, que durante o processo, realizou-se o processo da identificação dos elementos correspondentes, para possibilitar uma visão mais clara de quais são os atributos disponíveis, além de como estes se relacionam no contexto das diversas fontes de dados.

Figura 2 – Recorte do instrumento de mapeamento de estruturas de dados

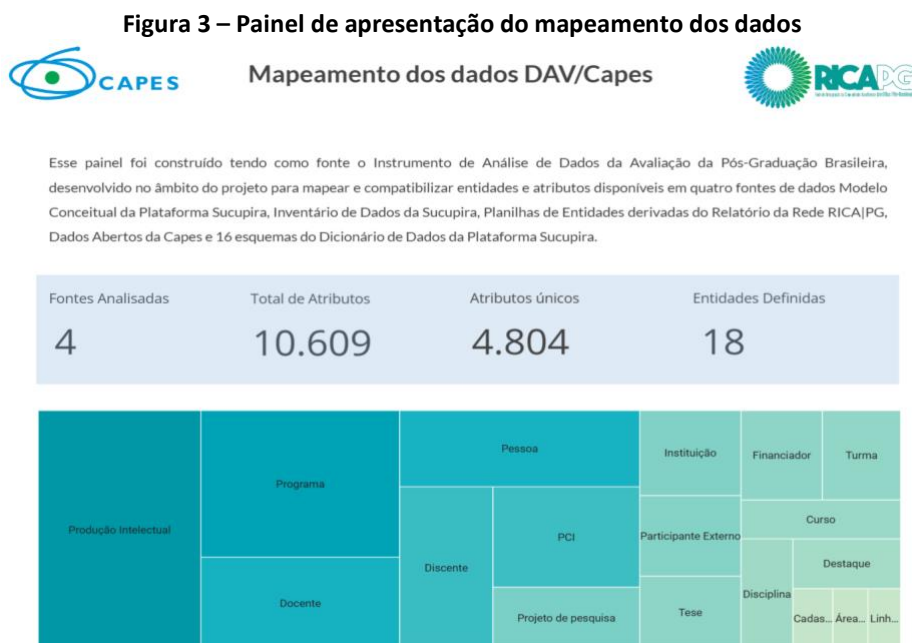
Entidade	Inventário	RICA-PG	Dados Abertos	Dicionário de Dados
9. Programa	Área básica		NM_AREA_BASICA	NM_AREA_BASICA
10. Produção Intelectual	ÁREA REPRESENTADA		DS_AREA_REPRESENTADA	
11. Área de concentração	Á(s) Áreas de concentração Ob...	NM_AREA_CONCENTRACAO	NM_AREA_CONCENTRACAO	NM_AREA_CONCENTRACAO
12. Disciplina	Á(s) Áreas de concentração Ob...	NM_AREA_CONCENTRACAO	NM_AREA_CONCENTRACAO	NM_AREA_CONCENTRACAO
13. Linha de pesquisa	Vínculo com Área de Concentra...		NM_AREA_CONCENTRACAO	NM_AREA_CONCENTRACAO

Fonte: Elaboração própria.

A inclusão das fontes, levou ao crescimento vertical do Instrumento, ao mesmo tempo, à medida que novos atributos e variáveis dessas fontes foram sendo acrescidos houve um crescimento horizontal, possibilitando uma visão macro do *status* dos dados. Tal processo foi inicialmente testado e validado para que fosse aplicado ao conjunto de atributos que compuseram a análise.

A partir dos dados estruturados no Instrumento de Análise de Dados da Avaliação da Pós-Graduação Brasileira, foi desenvolvido um ambiente de análise de dados utilizando a ferramenta de *Business Intelligence (BI) Looker Studio*, da *Google*, com o objetivo de apresentar os resultados alcançados. O ambiente foi designado como "Mapeamento dos Dados DAV/CAPES", que objetiva atuar como fonte de visualização dos dados das fontes (Inventário de Dados, RICA|PG, Dados Abertos, Dicionário de Dados) em suas estruturas originais e do mapeamento de dados das mesmas estruturas, realizado por meio do Instrumento de Análise de Dados da Avaliação da Pós-Graduação *stricto sensu* brasileira.

A Figura 3 apresenta o painel inicial do ambiente de BI desenvolvido, em que é possível identificar as diversas fontes analisadas, além de elementos como o quantitativo de atributos, e a sua divisão por meio das entidades disponíveis no contexto da avaliação.



Fonte: Elaboração própria.

Destaca-se que o ambiente de BI integra e apresenta cinco estruturas de dados, sendo as quatro fontes primárias e o Instrumento de Análise de Dados que as compatibiliza, totalizando 4.804 atributos únicos, exibidos com diferentes configurações em 20 páginas de painéis de visualização, para o apoio na análise quantitativa e qualitativa dos dados. Tais painéis foram utilizados para realizar análises da estrutura de dados mapeada, bem como possui potencial uso estratégico para visualizações e outros mapeamentos relacionados à Avaliação da pós-graduação brasileira.

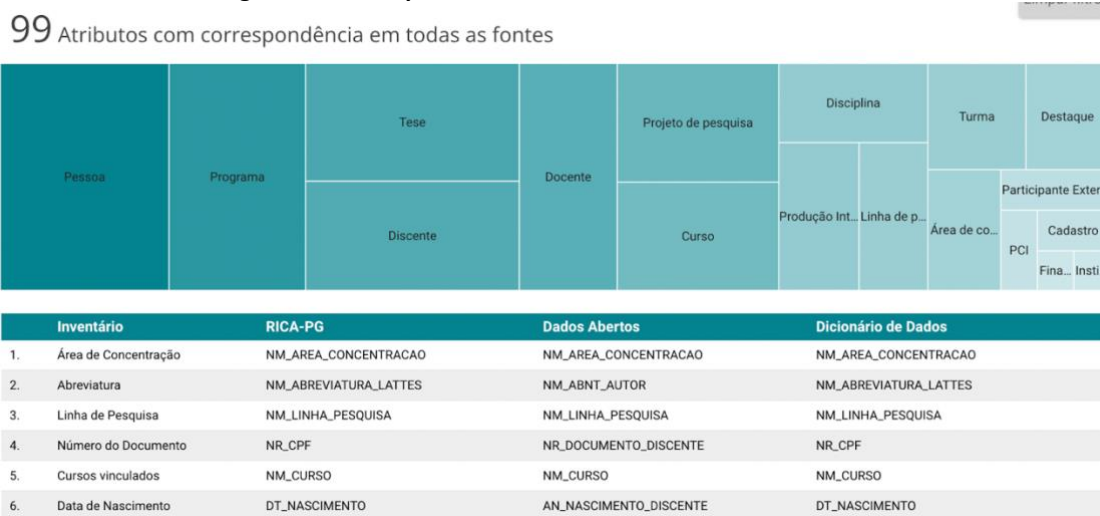
A partir do processo de mapeamento e da construção do instrumento e do ambiente de BI, verificou-se que, dos 4.804 atributos únicos, apenas 1.615 foram compatibilizados, ou seja, possuíam correspondência entre uma ou mais fontes. Dentre os atributos compatibilizados, a maioria refere-se à entidade Produção intelectual, seguida da entidade Programa de Pós-graduação.

Observou-se ainda que somente 99 atributos possuem correspondência em todas as estruturas analisadas, ou seja, dentre as estruturas das fontes de dados atuais da Capes, apenas 2% encontram correspondência completa. Esse percentual ressalta a

necessidade de integração entre as estruturas de dados. A Figura 4 apresenta esses atributos nas entidades mapeadas e alguns dos registros de atributos correspondentes nas 4 fontes analisadas.

Destaca-se que a compreensão e mapeamento dos atributos referem-se às estruturas de dados utilizadas para que os dados sejam armazenados nos sistemas de informação e tenham potencial de uso e reuso automatizados por processos de integração e interoperabilidade.

Figura 4 – Correspondência de atributos entre todas as fontes



Fonte: Elaboração própria.

Tal mapeamento é significativo, pois permitiu compreender a estrutura de dados utilizada para a representação de diferentes entidades e classes presentes na avaliação da pós-graduação *stricto sensu* brasileira no contexto da CAPES.

A partir dos dados mapeados, o estudo terá como continuidade o desenvolvimento da proposta de Arquitetura de Dados da Avaliação da Pós-Graduação Brasileira no contexto da CAPES, cuja aplicação poderá ser realizada nos sistemas de informação disponíveis, sobretudo na Plataforma Sucupira, além de servir de padrão para a comunicação de dados entre as Instituições e a CAPES, possibilitando que a integração de dados culmine em processos automatizados de alimentação de sistemas como a própria Plataforma Sucupira.

4 CONSIDERAÇÕES FINAIS

O presente estudo objetivou apresentar o mapeamento das estruturas de dados da avaliação da pós-graduação *stricto sensu* brasileira no contexto da CAPES, visando embasar a Arquitetura de Dados a ser desenvolvida como alicerce à integração de dados.

Esse processo possibilitará desenvolver uma Arquitetura de Dados da Avaliação da Pós-Graduação *stricto sensu* Brasileira no contexto da CAPES, por meio da qual as Instituições que ofertam Programas de Pós-Graduação poderão interagir, quer seja no processo de fornecimento ou de consumo dos dados de forma automatizada. Com isso, o uso e reuso dos dados poderá ser realizado com menor ruído de comunicação, gerando precisão, veracidade e velocidade ao processo.

A integração de dados de diferentes instituições e áreas do conhecimento pode se constituir em uma poderosa ferramenta de gestão, que possibilitará gerar análises, insights, apoiar a tomada de decisão, de forma eficiente e transparente, além de desenvolver pesquisas. Ademais, aponta-se o aspecto da transparência nos processos que envolvem a relação entre os Programas de Pós-Graduação e a CAPES sobretudo no que tange à avaliação.

Destaca-se ainda, que este trabalho contribui para o aprimoramento da integração entre os sistemas da CAPES e das instituições de Ensino Superior, e poderá apoiar a disseminação de boas práticas em integração, interoperabilidade e representação de dados. Este trabalho visa ainda otimizar os processos de avaliação da pós-graduação, além de contribuir para uma gestão mais eficiente e transparente dos dados da pós-graduação. Isto inclui os dados da produção científica, tendo as próprias instituições como fonte de proveniência dos dados, assegurando transparência, confiabilidade e eficácia ao processo de avaliação.

Portanto, o desenvolvimento do estudo contribui com o aprimoramento da gestão de dados da avaliação da pós-graduação *stricto sensu* brasileira pela CAPES, com potencial para impactar positivamente o processo de avaliação da pós-graduação *stricto sensu* brasileira.

REFERÊNCIAS

BRASIL. Ministério da Educação. **Programa de governança colaborativa de informações da pós-graduação (GoPG)**. [S. l.]: CAPES, [2024]. Disponível em:

<https://www.gov.br/capes/pt-br/aceso-a-informacao/acoes-e-programas/avaliacao/programa-de-governanca-colaborativa-de-informacoes-da-pos-graduacao-gopp>. Acesso em: 23 abr. 2024.

NÓBREGA, H. G. M.; TORINO, E.; VIDOTTI, S. A. B. G.; HALICKI, J. P. C.; OLIVEIRA, T. M. Ferramentas de apoio à integração de dados entre sistemas de informação no contexto da avaliação da pós-graduação stricto sensu pela CAPES. **Tendências da Pesquisa Brasileira em Ciência da Informação**, [S. l.], v. 16, n. 1, p. 1-21, dez. 2023. Disponível em: <https://revistas.ancib.org/index.php/tpbci/article/view/656>. Acesso em: 15 fev. 2024.

TORINO, E. **Arquitetura de dados no contexto da ciência da informação**. 2022. Tese (Doutorado em Ciência da Informação) – Universidade Estadual Paulista, Unesp, Marília, 2023. Disponível em: <https://repositorio.unesp.br/items/b2192b88-8362-488f-9b85-2c173eb66e48>. Acesso em: 15 fev. 2024.

TUPPER, C. D. **Data architecture: from zen to reality**. Amsterdam: Morgan Kaufmann, 2011.

YOON, V. Y.; AIKEN, P.; GUIMARAES, T. Applying a metadata framework to improve data quality. In: KHOSROW-POUR, M. (ed.). **Encyclopedia of information science and technology**. Hershey: Idea Group Reference, 2005. v. 1, p. 146-151.



[4.0 Internacional](https://creativecommons.org/licenses/by/4.0/)

Esta licença permite compartilhamento, remixe, adaptação e criação a partir do trabalho, mesmo para fins comerciais, desde que sejam atribuídos créditos ao(s) autor(es). Conteúdos elaborados por terceiros, citados e referenciados nesta obra não são cobertos pela licença.

Como citar este trabalho:

TORINO, Emanuelle; CONEGLIAN, Caio Saraiva; BRANDT, Mariana Baptista; OLIVEIRA, Talita Moreira de; VIDOTTI, Silvana Aparecida Borsetti Gregorio. Mapeamento de dados como arcabouço para a Arquitetura de Dados da Avaliação da Pós-Graduação Brasileira no contexto da CAPES. In: WORKSHOP DE INFORMAÇÃO DADOS E TECNOLOGIA, 8., 2025, Marília, SP. **Anais [...]**. Marília, SP: Universidade de Marília, 2025. DOI: <https://doi.org/10.22477/viii.widat.266>.

A Inteligência Artificial no horizonte da Ciência da Informação: dados de periódicos A1 entre os anos de 2015 a 2024

Artificial Intelligence on the Horizon of Information Science: data from A1 journals between 2015 and 2024

André Luiz Dias de França¹



<http://lattes.cnpq.br/5684942906411942>



<https://orcid.org/0000-0002-8278-3664>

Isaac Newton Cesarino da Nóbrega Alves²



<http://lattes.cnpq.br/4790445484081253>



<https://orcid.org/0009-0006-7727-3711>

Resumo

Trata-se de recorte da pesquisa intitulada “A Ciência da Informação: o que fizemos, em que lugar estamos e para onde vamos?” que investiga o escopo sobre o qual a Ciência da Informação no Brasil se desdobra entre 2015 e 2024. Possui natureza aplicada e quantitativa, analisou 1.540 publicações de quatro periódicos de extrato A1: Dados; Informação e Sociedade: estudos; Perspectivas em Ciência da Informação e; Transinformação. Como evidência ao que se apresenta, há um expressivo crescimento de menções sobre a Inteligência Artificial a partir de 2022, ano em que o ChatGPT é apresentado ao mundo e se populariza em todas as áreas da sociedade. Além disso, ficou constatada forte correlação da temática Inteligência Artificial com os seguintes termos: “direito autoral”, “plágio” e “propriedade intelectual”, no âmbito do Tesouro Brasileiro de Ciência da Informação, e; “traduções”, “gpt” e “Kaufman”, como termos amplos. Os achados indicam que o campo da CI vem demonstrando interesse no impacto da evolução da IA, quais são os outros interesses que orbitam tal temática no domínio da própria CI, sob o viés de seu tesouro e quais são os termos chaves mais amplos, dando conta de como o campo olha para além de seus limites.

Palavras-chave: Inteligência Artificial; Ciência da Informação; periódicos; tesouro.

Abstract

This study is an excerpt from the research titled “*Information Science: What Have We Done, Where Are We, and Where Are We Going?*”, which investigates the scope of Information Science in Brazil between 2015 and 2024. It is an applied and quantitative study that analyzed 1,540 publications from four A1-ranked journals: *Data, Informação e Sociedade: Estudos, Perspectivas em Ciência da Informação*, and *Transinformação*. The findings reveal a significant increase in mentions of Artificial Intelligence (AI) starting in 2022, the year ChatGPT was introduced and gained widespread popularity across various sectors of society. Additionally, a strong correlation was found between AI and the following terms within the *Tesouro Brasileiro de Ciência da Informação*: “copyright,” “plagiarism,” and “intellectual property.” Furthermore, broader terms associated with AI include “translations,” “GPT,” and “Kaufman.” The findings indicate that the field of Information Science has shown increasing interest in the impact of AI advancements, exploring the related topics within its domain through its

¹ Doutor, Universidade Federal da Paraíba (UFPB). Docente Universidade Federal da Paraíba (UFPB), João Pessoa, Paraíba, Brasil. Docente Colaborador no Programa de Pós-Graduação em Ciência da Informação (PPGCI) da Universidade Federal do Rio Grande do Norte (UFRN), Natal, Rio Grande do Norte, Brasil. andreluiz@ccta.ufpb.br.

² Mestre, Universidade Federal da Paraíba (UFPB). João Pessoa, Paraíba, Brasil. Técnico em Secretariado da Universidade Federal da Paraíba (UFPB). isaacnewtoncesarino77@gmail.com.

thesaurus, as well as identifying broader key terms that reflect how the field perceives AI beyond its immediate boundaries.

Keywords: Artificial Intelligence; Information Science; journals; *thesaurus*.

1 INTRODUÇÃO

A área da tecnologia com suas constantes inovações possibilitam que a população de maneira geral faça uso de novos produtos e serviços que perpassam na vida pessoal, profissional e acadêmica. Neste sentido, eleva-se nos últimos tempos o debate sobre a Inteligência Artificial (IA), especialmente o ChatGPT, por ser um modelo de IA que permite interações com os humanos por meio da linguagem natural.

Diante dessa realidade atual pela qual passa a sociedade da informação, julgamos ser pertinente direcionar nosso olhar no sentido de compreender como a Inteligência Artificial apresenta-se no campo da Ciência da Informação, especificamente, nos últimos anos sobre as pesquisas desenvolvidas e difundidas na CI.

1.1 Inteligência Artificial: Breve Consideração Conceitual

Atualmente, no mundo todo está em evidencia à chamada Inteligência Artificial (IA) e que para muitos, parece ser uma novidade fruto do avanço tecnológico recente. Porém a IA não é algo da atualidade, mas sim, uma tecnologia que surgiu anos atrás e que segundo Eysenck e Eysenck (2023, p. 6) “as verdadeiras origens da IA ocorreram muitos antes de 1955”. Ainda de acordo com os referidos autores, foi Ada Lovelace a “primeira programadora de computadores da história” e também responsável pela criação do primeiro algoritmo de máquina.

Outra contribuição relevante para a IA foi dada por Alan Turing quando faz a publicação de artigo no qual mostra a possibilidade da construção de máquinas capazes de realizar problemas utilizando 0s e 1s, levando o próprio pesquisador a desenvolver um aparelho de decifrar códigos.

Neste contexto, consideramos pertinente discorrer ainda que de maneira sucinta, sobre o significado de Inteligência Artificial (IA), termo cunhado por McCarthy *et al.* (2006) em 1955. Portanto, segundo Lopes, Santos e Pinheiro (2014, p. 1) trata-se de um termo que “constitui vários procedimentos computacionais cujas funções realizadas, caso um ser

humano as executasse, seriam consideradas inteligentes. O conceito é amplo e recebe tantas definições quanto os significados diferentes da palavra inteligência”. Na visão defendida por Santos (2021, p. 6), refere-se “[...] ao campo da ciência destinado a fornecer máquinas com a capacidade de realizar funções como lógica, raciocínio, planejamento, aprendizagem e percepção”.

Já para Sarlet, Sarlet e Bittar (2022, p. 7) a definição de Inteligência Artificial pode ser “uma capacidade de, por meio da tecnologia, se alcançar panoramas informacionais muito além dos outrora conhecidos, introduzindo novos critérios de tomada de decisão em função da absoluta incapacidade do cérebro humano”.

Ainda sobre o significado de IA, Gabriel (2022, p. 56) explica que:

[...] IA refere-se à capacidade das máquinas em imitar as funções da mente humana, fazendo com que os principais objetivos no desenvolvimento de IA, consequentemente, sejam dotar os computadores de características como: conhecimento, criatividade, raciocínio, solução de problemas complexos, percepção, aprendizagem, planejamento, comunicação em linguagem natural, habilidade de manipular e mover objetos, autonomia para tomada de decisão, entre outras habilidades que consideramos comportamento “inteligente”.

Desta forma, ao analisarmos os conceitos apresentados anteriormente neste estudo sobre IA, temos a nítida compreensão de que a Inteligência Artificial é a estrutura física (*hardware*) dotada de programas (*software*) que possibilitam à execução de tarefas idênticas ou superiores a capacidade humana e que estão em constantes transformações podendo, inclusive, receber a inclusão de neurônios gerados a partir de células-troncos, tornando-a ainda mais autônoma em diferentes habilidades e na tomada de decisão nos mais variados campos científicos, não se limitando a apenas aprender e emitir respostas numa interação com humanos.

1.2 Objetivos e Aplicações da IA

O surgimento de uma determinada tecnologia está voltado para a consecução de algo em favor da humanidade, portanto, a Inteligência Artificial apresenta dois objetivos, sendo um tecnológico e outro científico. No âmbito tecnológico, objetiva “fazer coisas úteis (às vezes empregando métodos muitos diferentes dos utilizados pela mente)” afirma Boden (2020, p. 14). No tocante ao aspecto científico, a referida autora ainda expõe que a Inteligência Artificial

busca “usar conceitos e modelos de IA para ajudar a responder perguntas sobre os seres humanos e outros seres vivos”.

Com o passar dos anos, a Inteligência Artificial passou e ainda passa nos dias atuais por inúmeras transformações, sendo esta constatação amparada no que dizem Faceli *et al.* (2023, p. 1) quando afirmam que “desde os anos 1970, houve uma expansão do uso da Inteligência Artificial (IA) para a solução de problemas reais”.

Hoje, a IA atua em muitas áreas do conhecimento humano e suas aplicações podem ser perceptíveis desde o simples reconhecimento facial ou de voz humana, passando por complexos procedimentos de saúde em pacientes de países avançados e com aparato tecnológico significativamente avançado (Faceli *et al.*, 2023). Dentre as muitas aplicações da IA, destacam-se:

- a) Matemática: demonstração de teoremas, resolução simbólica de equações, aproximação de funções complexas etc.;
- b) Pesquisa operacional: otimização e busca heurística em geral;
- c) Jogos: xadrez, damas, jogo etc.;
- d) Processamento de linguagem natural: tradução automática, verificadores ortográficos e sintáticos, interfaces para BDs etc.;
- e) Sistemas tutores: modelagem de aprendizados, escolha de estratégias pedagógicas etc.;
- f) Percepção: visão, tato, audição, olfato, paladar e outras;
- g) Robótica: visão, navegação, controle etc.;
- h) Sistemas especialistas: atividades que exigem conhecimento especializado, e não formalizado; extração de conhecimentos; regras ou informações em tarefas como diagnósticos, previsão, monitoramento, análises, planejamentos e projeto; e em áreas como Medicina, Finanças, Engenharia, Artes, Computação e outras. (Lopes; Santos; Pinheiro, 2014, p. 6).

Ainda no contexto atual, ressaltamos que uma das ferramentas de IA muito difundidas e com a possibilidade de uso pelas pessoas na sociedade contemporânea é denominada de ChatGPT e que ganhou notoriedade diante das muitas matérias jornalísticas que instigou a curiosidade, utilização e reação por parte de muitos cidadãos ao redor do mundo.

1.3 IA na Ciência da Informação: Relações Interdisciplinares

Desde o seu surgimento, a Ciência da Informação (CI) busca dialogar com outros campos científicos o que a caracteriza como interdisciplinar. Nesse sentido, alguns

profissionais atuaram de maneira efetiva para essa realidade, como bem destaca Saracevic (1996, p. 48) ao dizer:

A interdisciplinaridade foi introduzida na CI pela própria variedade da formação de todas as pessoas que se ocuparam com os problemas descritos. Entre os pioneiros havia engenheiros, bibliotecários, químicos, linguistas, filósofos, psicólogos, matemáticos, cientistas da computação, homens de negócios e outros vindos de diferentes profissões ou ciências. Certamente, nem todas as disciplinas presentes na formação dessas pessoas tiveram uma contribuição igualmente relevante, mas essa multiplicidade foi responsável pela introdução e permanência do objetivo interdisciplinar na CI.

Outra constatação sobre a interdisciplinaridade na CI e as áreas com relações de maior proximidade no início do nascimento deste campo do saber humano podem ser também conhecidas a partir da fala de Oliveira (2011, p. 20), quando expõe:

A participação de outros campos do conhecimento na Ciência da Informação permanece em função da complexidade dos problemas a serem equacionados pela área, o que exige a contribuição de diferentes profissionais e/ ou pesquisadores. Dentre as disciplinas com as quais a Ciência da Informação tem trabalhado distinguem-se: Biblioteconomia, Ciência da Computação, Comunicação Social, Administração, Linguística, Psicologia, Lógica, Matemática, Filosofia/ Epistemologia.

Ao analisarmos as falas anteriormente apresentadas, percebemos a presença primordial da Ciência da Computação como uma das disciplinas pioneiras envolvidas com a CI. Mooers (1951) ao tratar da temática da recuperação da informação aborda os sistemas ou máquinas utilizadas na execução de tarefas, deixando evidenciada mais uma vez a relevância da tecnologia para a Ciência da Informação. Diante disso, vale salientar que “as relações interdisciplinares estão evoluindo em direção a um nível diferente de cooperação intelectual” de acordo com Saracevic (1996, p. 50).

Por conseguinte, na literatura disponibilizada sobre a interdisciplinaridade na CI, tanto os autores aqui enfatizados, bem como outros reforçam a contribuição da área da tecnologia para a Ciência da Informação desde o passado até os dias atuais. Desta forma, a presença da tecnologia na CI vem desde sua criação e esta compreensão ampara-se em Oliveira (2011) ao relatar que Vanevar Bush, na década de 1945, desejava disseminar a produção dos conhecimentos do período posterior a Segunda Guerra Mundial a partir de tecnologias da informação.

Ainda de acordo com a referida autora, “ele chegou a propor uma máquina com capacidade de ‘associar ideias’, que duplicaria os ‘processos mentais artificialmente’”

(Oliveira, 2011, p. 12). Neste caso, a autora está se referido ao uso do computador ao falar de máquina para executar processos mentais de maneira artificial, logo, percebemos a ainda que de maneira indireta, a citação sobre a participação da Inteligência Artificial (IA) nas atividades de processamento de dados, já que a mesma opera por meio das redes neurais artificiais (RNAs).

Diante do contexto da IA na Ciência da Informação, Saracevic (1996, p. 50) esclarece que:

Uma das áreas chaves de interesse para ambas, ciência da computação e CI, é a Inteligência Artificial (IA). Embora a IA possa ser discutida por si só, prefiro discuti-la na perspectiva da ciência cognitiva para ilustrar os interesses compartilhados acerca dos processos da mente. Enfim, a fonte e o alvo da CI tem sido a informação da e para a mente.

Portanto, a IA pode embasar e também contribuir efetivamente com estudos no campo da CI, principalmente no que tange aos processos mentais da execução das operações, o que nos leva a lembrar da contribuição dos paradigmas da Ciência da Ciência, de autoria de Rafael Capurro (2003), ao trazer para o centro dos debates a questão do paradigma cognitivo.

Na visão de Saracevic (1996, p. 52), a Inteligência Artificial é classificada em IA forte e IA fraca, sendo que a “IA forte, com suas teorias da inteligência, mente e pensamento conectadas com as máquinas, seja controversa, causando numerosos desafios, especialmente entre filósofos”. O autor ainda esclarece que:

Tanto a IA fraca quanto a forte tem interesse direto para a CI. A IA fraca é fonte de muitas das inovações nos sistemas de informação, tais como sistemas inteligentes, hipertextos, bases de conhecimento, interfaces inteligentes e as questões sobre a interação homem-computador - todas elas de interesse para a CI e para as quais ela pode contribuir diretamente. A IA forte é a fonte do modelo teórico da cognição, no qual a informação, enquanto fenômeno, desempenha o mais importante papel. Portanto, esse modelo pode também contribuir para a pesquisa básica em CI. (Saracevic, 1996, p. 52).

Logo, a Inteligência Artificial diante das inovações percebidas na atualidade, pode contribuir e instigar novas pesquisas na Ciência da Informação, buscando dirimir problemas presentes na sociedade, além, evidentemente, de interligar e abranger ainda mais o campo interdisciplinar que detém junto a CI.

2 PROCEDIMENTOS METODOLÓGICOS

Trata-se de pesquisa exploratória, de abordagem quantitativa, que utilizou um *corpus* formado por 1.540 publicações obtidas nos *sites* institucionais dos periódicos de extrato A1³ Dados (UERJ); Informação & Sociedade (UFPB); Perspectivas em Ciência da Informação (UFMG) e; Transinformação (PUC-SP). Das publicações, originalmente em *Portable Document Format* (PDF), foram extraídas as partes textuais, descartando-se imagens e quaisquer outros elementos irrelevantes para a Mineração de Texto. Este tratamento envolveu 5 etapas, quais foram: Coleta, já mencionada; Pré-processamento; Indexação; Mineração e; Análise (Aranha e Passos, 2006, p. 4), com intuito da Descoberta de Conhecimento em Textos (*Knowledge Discovery in Texts* – KDT).

Para Feldman e Dagan (1995, p. 112, tradução nossa), por mais que a *Knowledge Discovery in Databases* (KDD) se volte “[...] para bancos de dados estruturados, é evidente que esse paradigma também é necessário para lidar com a enorme quantidade de informações disponíveis apenas em forma textual não estruturada”. Para os autores, com o volume crescente de informações na era da informação em mídias eletrônicas, a KDD torna-se um novo paradigma no que tange o estudo sobre grandes *corpus*. (Feldman; Dagan, 1995). Nesse âmbito, insere-se a KDT como uma abordagem viável no escopo da KDD.

A manipulação do *corpus* foi realizada através da Linguagem R diante do seu potencial na abordagem quantitativa e da sua vasta documentação referencial, sendo ainda um ambiente em que seu código aberto fomenta a colaboração e o aumento de soluções para os mais variados problemas de pesquisa científica. Nesse contexto, consultamos o Tesouro Brasileiro de Ciência da Informação (TBCI) (Pinheiro; Ferrez, 2014) para obtermos os termos associados ao campo da CI e assim delimitar os resultados, em um primeiro momento. Como adiantado, também usando a Linguagem R, extraímos de cada PDF a camada textual resultando em um novo arquivo (TXT). Esse novo documento compôs a massa de igual tamanho (1.540) e que alimentou a etapa do pré-processamento. Nesse ponto, descartamos os caracteres especiais, e, na sequência, homogeneizado em letras minúsculas, excluindo-se ainda elementos numerais, pontuações e espaços extras.

³ Qualis A1 listados na Plataforma Sucupira¹ em ao menos um dos dois últimos quadriênios (2013-2016 e 2017-2020).

No intuito de evitar que houvesse um duplo achado, por exemplo, para “inovação” em “inovação” e “inovação tecnológica”, no processo de tokenização, substituímos os espaços por “_”, de modo que o segundo termo passou a ser “inovação_tecnológica”. Com a etapa de pré-processamento concluída, “atravessamos” o nosso *corpus* pela “malha” composta por termos do TBCI (Pinheiro; Ferrez, 2014), de sorte que tivéssemos um subconjunto de elementos compondo uma matriz de frequências. Em um segundo momento, realizamos procedimentos para obtermos quais os termos do TBCI (Pinheiro; Ferrez, 2014) se correlacionavam com o termo “Inteligência Artificial”, destaque para esse trabalho, bem como para termos além daquele tesouro (sem filtro terminológico algum), de modo que pudéssemos vislumbrar para quais outros horizontes a CI poderia ampliar seu escopo de abordagem.

A despeito dessa abordagem, utilizamos a medida estatística Correlação de Pearson. Trata-se de uma medida linear que expressa a relação entre duas variáveis. Essa abordagem lançou mão da função **findAssocs()** embutida na biblioteca **tm** (Text Mining) na Linguagem R. Sabemos que isso pressupõe a distribuição normal dos dados, o que não observamos em nossa amostra. No entanto, realizamos uma validação em que comprovamos para os anos de 2023 e 2024, que a amostra é suficientemente grande conforme a Teorema Central do Limite (TCL), a distribuição é próxima da normal obtidas por *Bootstrap*, o tamanho amostral está adequado (Bonett; Wright, 2000) e as correlações de Pearson são fortes e consistentes durante os anos.

Nesse âmbito, sentimos confiança nos resultados obtidos, sobretudo porque parecem fazer sentido quando olhamos o cenário em que se apresentaram como poderemos observar na próxima seção e bem como a preocupação da academia com os impactos da IA.

3 RESULTADOS

O TBCI (2014) é uma obra que abarca 1.813 termos que foram considerados relevantes para cobrir o campo científico da CI. Criado em 2014, esse tesouro inclui a IA de modo que indica a preocupação da CI em manter em seu horizonte de pesquisa, os impactos de sua atuação. Para este trabalho, observamos que tal termo figurou [apenas] na posição 189 quando mineramos aqueles mais abundantes nas publicações analisadas dos últimos dez anos. Sua presença nesse extrato demonstra a atenção do campo com o desenvolvimento da

IA, contudo, no recorte de uma década, e uma massa documental de 1.540 publicações, parece tornar-se menos importante, diante de outros centrais, como vemos na Figura 14.

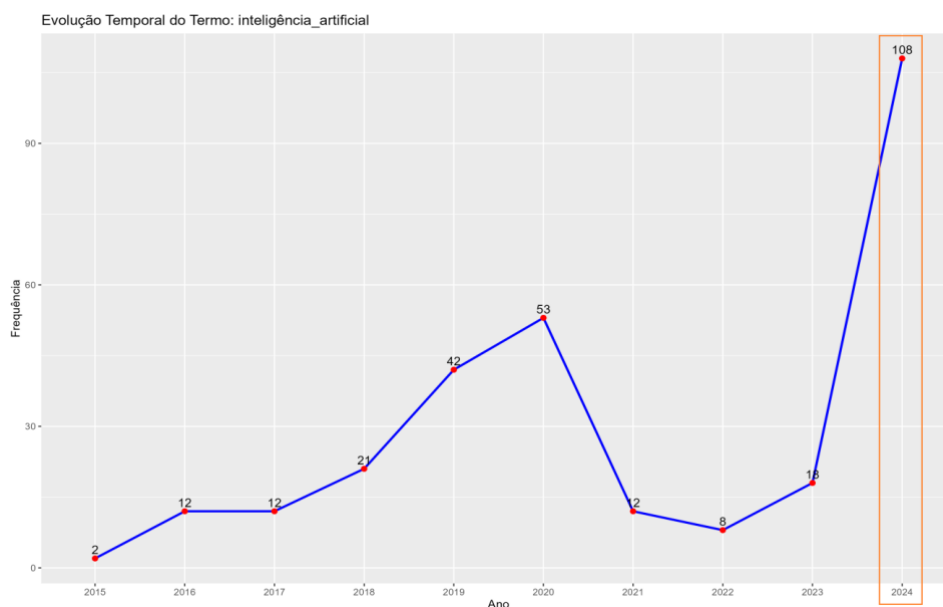
Figura 1 – 250 termos mais frequentes



Fonte: Elaboração própria.

Citados 288 vezes, de longe, não representa relevância, em nossa opinião, quando olhamos a década como um todo, mas surge no horizonte da CI quando salta de 18 menções em 2023 para 108 em 2024, colocando-o em 48º lugar dentre todos do TBCI (2014), quando observamos a evolução temporal (Gráfico 1).

Gráfico 1 – Evolução Temporal – 2015-2024



Fonte: Elaboração própria.

⁴ A figura está deliberadamente cortada para destacar apenas o quão o termo Inteligência Artificial está deslocada dos termos mais frequentes.

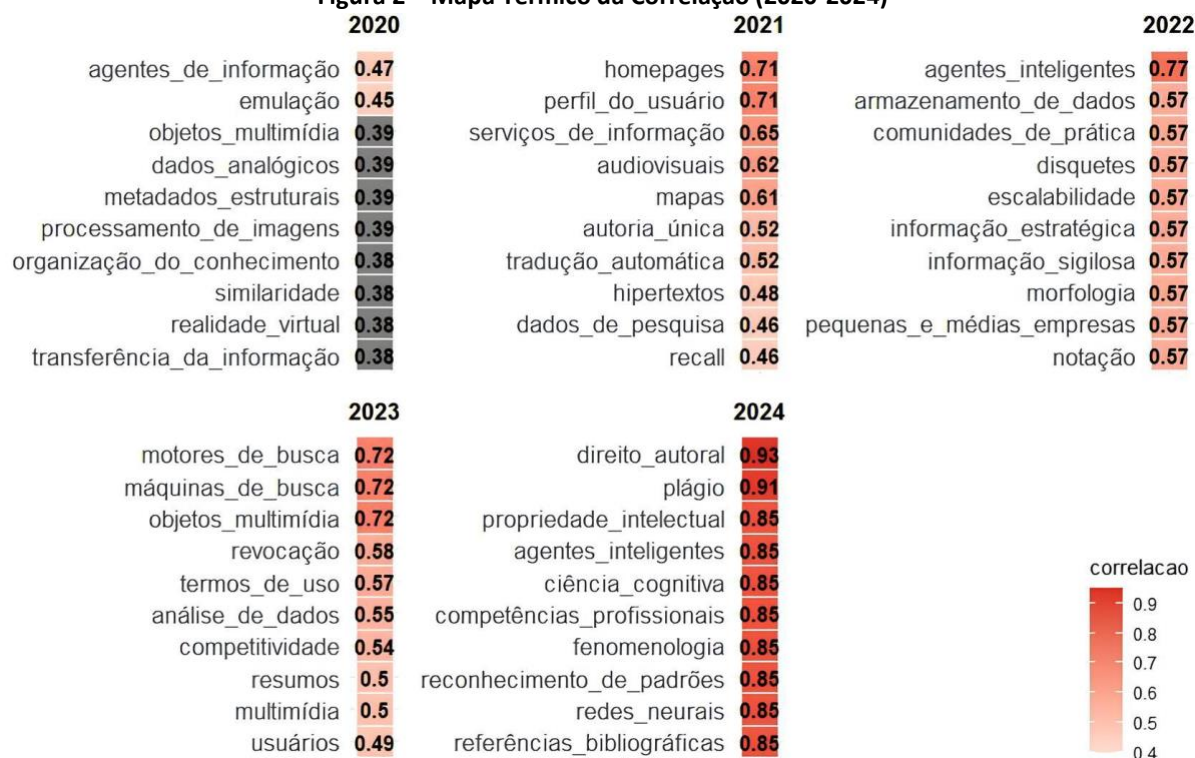
Em 30 de novembro de 2022, a OpenAI (2025, tradução nossa) apresenta ao mundo sua Inteligência Artificial. A empresa tem por missão, “[...] garantir que a Inteligência Artificial geral beneficie toda a humanidade.” Sua ferramenta, o ChatGPT (2025), é “[...] um modelo avançado de Inteligência Artificial projetado para gerar respostas em linguagem natural, simulando conversas humanas com precisão e flexibilidade.” Veríamos, nos anos seguintes, algo disruptivo em termos de tecnologia e logo impactaria o mundo em que vivemos, incluindo a academia, a pesquisa e a ciência. Essa onda de impacto pode ser notada no gráfico anterior no ano de 2024.

A CI, dada sua indissociabilidade como o imperativo tecnológico, sentiu, naturalmente, o abalo que aquele lançamento proporcionou entre suas e seus pesquisadores. Ao considerarmos o tempo em que a comunidade científica se aproxima de seu problema de pesquisa, explora-o, reflete, submete seus achados e tem aceitos seus escritos para eventual publicação, podemos considerar que o fim de 2022 e todo 2023, foi o tempo em que a CI cristalizou sua preocupação com o avanço do assunto.

Em uma segunda etapa e para ampliar o interesse do campo em torno do tema, procuramos descobrir a que demais termos do TBCI (2014), o assunto “Inteligência Artificial” estaria correlacionado. Do gráfico anterior – destacamos a frequência 108 (2024) – obtivemos as correlações a seguir (Figura 2).

Ao considerarmos que uma correlação moderada o valor de 0,5 e uma perfeita, em 1,0, observamos que, entre 2021 e 2024, a comunidade refletiu sobre questões distintas. Em 2020, com correlações mais fracas, o foco estava nas bases conceituais da IA, como conteúdos multimídias e simulações além da preocupação com a mediação da informação. Em 2021, as correlações parecem se aproximar das interfaces e dos usuários. Já em 2022, a visão estratégica fomenta o debate do uso prático e organizacional da IA. No ano de 2023, fortemente vemos a IA associada à busca, análise de dados e competitividade, consolidando-a como interface de acesso e uso da informação. Em 2024, quando as correlações estão muito fortes, diante de seu rápido amadurecimento no tocante ao uso pelas pessoas, a ciência parece se preocupar com questões éticas e jurídicas da experiência do ser humano em introduzir a IA em seus afazeres. O cerne nesse ano se resume, podemos concluir, no fazer o bem e fazer o certo.

Figura 2 – Mapa Térmico da Correlação (2020-2024)



Fonte: Elaboração própria.

Em suma, podemos assumir que a evolução se deu assim: a) 2020 - Fundamentos da atual IA; b) 2021 - Usuário e aplicação prática; c) 2022 - Pensamento estratégico organizacional; d) 2023 - Recuperação, dados e competitividade e; e) 2024 - Uso moral e ético da IA. Mas, para 2024, ano em que as correlações se apresentaram fortíssimas – para efeito de comparação, a menor delas (0.85) é superior a maior (0.77) dos anos anteriores – indicando que a academia está, atualmente, preocupada com os impactos morais e éticos do uso da IA.

Portanto, foi intuito nosso expandir a busca e ir além do TBCI, ao buscarmos saber como se daria a correlação da Inteligência Artificial para além dos termos do campo da CI. Como isso, encontramos o seguinte: direito_autoral – 0.93; gpt – 0.92; plágio – 0.91; chat 0.9 e; kaufman – 0.9. Seja no escopo do TBCI (2014) ou além de seus limites, “direito autoral” se mantém correlacionado à “Inteligência Artificial”, acompanhado de “gpt”, o que seguramente remete à criação da OpenAI, e novamente de “plágio”, sacramentando a preocupação com as questões éticas de seu manuseio e uso de seus resultados, por assim dizer. Por fim, temos “kaufman”, que associamos a pesquisadora e autora Dora Kaufman, especialista em Inteligência Artificial, dando conta que a professora possui elevada relevância, também, entre aqueles que publicaram trabalhos na área.

4 CONSIDERAÇÕES FINAIS

Consideramos que o trabalho aqui exploratório possui potencial ao trazer à luz que as pesquisadoras e pesquisadores que publicaram seus trabalhos naqueles quatro períodos de relevância no País, e para a Ciência da Informação, evidenciou o crescimento no interesse em se discutir a Inteligência Artificial bem como seus impactos fortemente nos derradeiros anos. Também é interessante destacar os assuntos nucleares que circundaram a discussão: de que forma conseguiremos sustenta uma inovação tecnológica com o uso responsável, ético e legal? Já ao final, quando observamos para além do domínio da CI, observamos que o lançamento disruptivo do ChatGPT ao fim de 2022, foi central na discussão, mantendo a questão do direito autoral e do plágio como assuntos correlatos.

No tocante às limitações estão no campo editorial, uma vez que não temos espaço aqui para ampliar o debate, e na abordagem do uso do coeficiente de correlação de Pearson, uma vez que ele mede tão somente o quanto duas variáveis apareceram juntas frequentemente. São indícios de uma relação, mas não é conclusivo. Ainda assim, cremos o suficiente para afirmar que a academia está interessada fortemente na discussão que envolva ética e IA.

Com indicações de trabalhos futuros, analisaremos em que cenários, os termos aqui em destaque, se apresentam e que outras ramificações terminológicas eles desencadearam. Para além do escopo de publicações em periódicos, continua em andamento – no âmbito da pesquisa ampla “A Ciência da Informação: o que fizemos, em que lugar estamos e para onde vamos?” – a coleta das teses e dissertações no mesmo recorte temporal (2015 a 2024) de modo que poderemos, também, observar sobre o que se debruçaram os candidatos a mestres e doutores em CI no Brasil e se e sob que correlações a IA se revelou.

REFERÊNCIAS

ARANHA, Christian; PASSOS, Emmanuel. A tecnologia de mineração de textos. **Revista Eletrônica de Sistemas de Informação (RESI)**, [Curitiba], v. 5, n. 2, p. 1-8, 2006. Disponível em: <http://www.periodicosibepes.org.br/index.php/reinfo/article/view/171>. Acesso em: 9 mar. 2025.

BODEN, Margaret A. **Inteligência Artificial**: uma brevíssima introdução. São Paulo: Editora Unesp, 2020.

BONETT, Douglas G.; WRIGHT, Thomas A. Sample size requirements for estimating pearson, kendall and spearman correlations. **Psychometrika**, [S. l.], v. 65, n. 1, p. 23-28, 2000. Disponível em: <https://link.springer.com/article/10.1007/BF02294183>. Acesso em: 9 mar. 2025.

CAPURRO, Rafael. Epistemologia e ciência da informação. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 5., 2003, Belo Horizonte. **Anais [...]**. Belo Horizonte: UFMG, 2003.

CHATGPT. **ChatGPT FAQ**. 2025. Disponível em: <https://chatgpt.com.br/faq/>. Acesso em: 27 fev. 2025.

EYSENCK, Michael W.; EYSENCK, Christine. **Inteligência Artificial x humanos**: o que a ciência cognitiva nos ensina ao colocar frente a frente a mente humana e a IA. Porto Alegre: Artmed, 2023.

FACELI, Katti *et al.* **Inteligência Artificial**: uma abordagem de aprendizado de máquina. Rio de Janeiro: LTC, 2023.

FELDMAN, Ronen; DAGAN, Ido. Knowledge Discovery in Textual Databases (KDT). In: INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING – KDD, 1., 1995, Montreal. **Proceedings [...]**. [S. l.]: AAAI Press, 1995. p. 112-117. Disponível em: <http://www.aaai.org/Papers/KDD/1995/KDD95-012.pdf>. Acesso em: 09 mar. 2025.

GABRIEL, Martha. **Inteligência Artificial**: do zero ao metaverso. Barueri, São Paulo: Atlas, 2022.

LOPES, Isaia Lima; SANTOS, Flávia Aparecida Oliveira; PINHEIRO, Carlos Alberto Murari. **Inteligência Artificial**. Rio de Janeiro: Elsevier, 2014.

MCCARTHY, John *et al.* A proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. **AI Magazine**, [S. l.], v. 27, n. 4, p. 12, 2006. DOI: <https://10.1609/aimag.v27i4.1904>. Disponível em: <https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1904>. Acesso em: 9 jan. 2025.

OLIVEIRA, Marlene de. **Ciência da Informação e Biblioteconomia**: novos conteúdos e espaços de atuação. Belo Horizonte: Editora UFMG, 2011.

OPENAI. **About**. 2025. Disponível em: <https://openai.com/about/>. Acesso em: 27 fev. 2025.

SANTOS, Marcelo Henrique dos. **Introdução à Inteligência Artificial**. Londrina: Editora e Distribuidora Educacional S. A., 2021.

SARACEVIC, Tefko. Ciência da Informação: origem, evolução e relações. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 1, n. 1, p. 41-62, jan./jun. 1996. Disponível em: <https://brapci.inf.br/index.php/res/v/37415>. Acesso em: 23 fev. 2025.

SARLET, Ingo Wolfgang; SARLET, Gabrielle B. Sales; BITTAR, Eduardo C. B. **Inteligência Artificial, proteção de dados pessoais e responsabilidade na era digital**. São Paulo: Expressa Jur, 2022.

PINHEIRO, Lena Vânia Ribeiro; FERREZ, Helena Dodd (org.). **Tesouro Brasileiro de Ciência da Informação (TBCI)**. Rio de Janeiro; Brasília: Ibict. 2014.



[4.0 Internacional](https://creativecommons.org/licenses/by/4.0/)

Esta licença permite compartilhamento, remixe, adaptação e criação a partir do trabalho, mesmo para fins comerciais, desde que sejam atribuídos créditos ao(s) autor(es). Conteúdos elaborados por terceiros, citados e referenciados nesta obra não são cobertos pela licença.

Como citar este trabalho:

FRANÇA, André Luiz Dias de; ALVES, Isaac Newton Cesarino da Nóbrega. A Inteligência Artificial no horizonte da Ciência da Informação: dados de periódicos A1 entre os anos de 2015 a 2024. *In: WORKSHOP DE INFORMAÇÃO DADOS E TECNOLOGIA*, 8., 2025, Marília, SP. **Anais [...]**. Marília, SP: Universidade de Marília, 2025. DOI: <https://doi.org/10.22477/viii.widat.256>.

Desenvolvimento de coleção digital voltada à governança em biblioteca digital*Development of a digital collection for governance in a digital library***Tiago Emmanuel Nunes Braga¹**<http://lattes.cnpq.br/8376134230259399><https://orcid.org/0000-0001-6332-7965>**Milton Shintaku²**<http://lattes.cnpq.br/8605833104600600><https://orcid.org/0000-0002-6476-4953>**Cristiane Henrique Santos³**<http://lattes.cnpq.br/6194941001104710><https://orcid.org/0009-0007-4437-1582>**Maison Roberto Mendonça Gonçalves⁴**<http://lattes.cnpq.br/4801149802103690><https://orcid.org/0000-0003-4056-6917>**Mirele Carolina Souza Ferreira Costa⁵**<http://lattes.cnpq.br/8547303047227327><https://orcid.org/0000-0002-1337-4672>**Resumo**

Manter a documentação de governança acessível na *web* faz parte das próprias orientações para melhor gestão. Por isso, o presente estudo apresenta o relato de experiência na implementação da Biblioteca Digital de Governança, do Instituto Brasileiro de Informação em Ciência e Tecnologia, com caráter aplicado. Para tanto, foi utilizada a tecnologia Omeka, com organização das coleções por tipologia documental e padrão Dublin Core para sua descrição. Como funcionalidade adicional, foi criada uma linha do tempo baseada em documentação

¹ Doutor em Ciência da Informação, Universidade de Brasília (UnB), Brasília, DF, Brasil. Diretor, Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), Brasília, DF, Brasil. tiagobraga@ibict.br.

² Doutor em Ciência da Informação, Universidade de Brasília (UnB), Brasília, DF, Brasil. Coordenador de Tecnologias para Informação (Cotec), Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), Brasília, DF, Brasil. shintaku@ibict.br.

³ Especialista em Língua Portuguesa e Literatura no Contexto Educacional, Centro Universitário Internacional (UNINTER), Curitiba, PR, Brasil. Revisora, Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), Brasília, DF, Brasil. cristianehenrique@ibict.br.

⁴ Bacharel em Biblioteconomia, Universidade Federal do Rio Grande (FURG), Rio Grande, RS, Brasil. Pesquisador bolsista, Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), Brasília, DF, Brasil. maisongoncalves@ibict.br.

⁵ Mestre em Informática, Universidade de Brasília (UnB), Brasília, DF, Brasil. Pesquisadora bolsista, Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), Brasília, DF, Brasil. mirelecosta@ibict.br.

relacionada ao Instituto. Com isso, apresentou-se uma opção para criação de bibliotecas digitais mais simples e eficientes para disseminação da documentação de governança do Instituto.

Palavras-chave: governança; biblioteca digital; coleção digital; Instituto Brasileiro de Informação em Ciência e Tecnologia; Omeka.

Abstract

Keeping governance documentation accessible on the web is part of the guidelines for better management. Therefore, this study presents an experience report on the implementation of the Digital Governance Library of the Instituto Brasileiro de Informação em Ciência e Tecnologia, with an applied approach. For this purpose, the Omeka technology was used, organizing collections by document typology and adopting the Dublin Core standard for description. As an additional feature, a timeline was created based on documentation related to the institute. Thus, the initiative provided a viable option for creating simpler and more efficient digital libraries to disseminate the institute's governance documentation.

Keywords: governance; digital library; digital collection; Brazilian Institute of Science and Technology Information; Omeka.

1 INTRODUÇÃO

Algo que é comum a todas as organizações, instituições e governo é requerer gestão guiada por um conjunto de orientações, comumente denominadas de governança, para que possam atuar de forma eficiente. Pela importância da governança dentro do governo, a Controladoria Geral da União (CGU), em seu *site*, dedica uma página a esse tema⁶, na medida em que está presente no Decreto nº 9.203, de 22 de novembro de 2017, que dispõe sobre a governança da administração pública federal direta, autárquica e fundacional.

Nesse mesmo caminho, o Tribunal de Contas da União (TCU), nos 10 passos para a boa governança (Brasil, 2021), descreve a necessidade de modelo de governança adequado a sua estrutura, íntegra e com transparência, de forma a possibilitar auditorias e prestações de contas. Assim, tem-se a necessidade do uso de sistema de informação para gerenciar a documentação do próprio modelo de governança, entre outros documentos, que fazem parte da memória técnica, de forma a dar acesso a todos os colaboradores, e promover a transparência ativa dessa documentação.

Cabe salientar que, no caso dos órgãos de governo, portarias, decretos e planos, por exemplo, são documentos importantes para a governança, pois orientam o seu funcionamento, sendo útil em processos de tomada de decisão ou mesmo nas atividades

⁶ Disponível em: <https://www.gov.br/cgu/pt-br/centrais-de-conteudo/campanhas/integridade-publica/governanca#:~:text=A%20governan%C3%A7a%20pode%20ser,com%20foco%20em%20objetivos%20coletivos>. Acesso em: 03 fev. 2025.

cotidianas. Por isso, essa documentação deve estar disponível a todos os colaboradores do órgão, de forma fácil, preferencialmente em sistema de informação, pois como salienta Monteiro, Carelli e Pickler (2008), essa documentação está sempre ameaçada do esquecimento.

Nesse contexto, o presente trabalho apresenta um estudo voltado a relatar a criação da Biblioteca Digital de Governança (BDG), do Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), e a customização da ferramenta livre Omeka para a criação da BDG, com organização por coleção e representação por metadados adequado para cada tipologia, a fim de contribuir com estudos sobre bibliotecas digitais e o uso de tecnologias variadas.

1.1 Governança

Governança, como é entendido atualmente, é um termo relativamente novo no léxico brasileiro, com raízes no verbo governar, seguido pelo sufixo -ança, formador de substantivo oriundo de verbo, que dá a ideia de ação ou o seu resultado. Sendo assim, governança, de forma básica, é o processo de governar ou os seus resultados. Governar, por sua vez, é estabelecer rumos e formas de alcançá-los, com legitimidade, visto que pela sua etimologia, vem do grego, para nomear o piloto de um barco.

A partir da conceituação etimológica, a governança passou a ser mais utilizada na esfera pública, tanto que Peters (2013, p. 28) diz que o termo governança descreve “como o setor público e outras instituições gerenciam a si mesmos bem como suas relações com a sociedade mais ampla”. Este ponto tornou-se tão importante, que foi promulgado um decreto dispondo sobre a política de governança da administração pública federal direta, autárquica e fundacional, o Decreto nº 9.203, de 22 de novembro de 2017.

Para esse decreto (Brasil, 2017), os princípios da governança são: (i) capacidade de resposta; (ii) integridade; (iii) confiabilidade; (iv) melhoria regulatória; (v) prestação de contas e responsabilidade; e (vi) transparência. Assim, a implementação de uma boa governança requer ações que atendam os seus princípios, nos quais pode-se extrair da lista a regulamentação, por meio de instrumentos oficiais, e a disseminação dessas ações de forma ampla, atendendo a todos os interessados.

Alinhando a governança à gestão da informação, para que seja efetivada a informação sobre governança deve circular, sendo de fácil acesso. Assim, nada mais efetivo

que um sistema de informação disponível na *web*, como as bibliotecas digitais, que possibilitam a gestão do acervo e formas diversas de recuperação da documentação.

2 PROCEDIMENTOS METODOLÓGICOS

O presente estudo é um relato de experiência da implementação da BDG do Ibict. Assim, para cumprir os objetivos do estudo, foram definidas etapas conforme o Quadro 1. Com isso, partiu-se do problema de pesquisa sintetizado na seguinte pergunta: “Como disseminar a documentação de governança do Ibict por meio de uma biblioteca digital?”. A seleção do sistema de informação Biblioteca Digital deu-se devido à ampla experiência do Instituto na gestão de documentos de primeira fonte.

Quadro 1 – Etapa do estudo

Nº	Etapa	Descrição
1	Análise das necessidades de gestão da documentação de governança.	Conversa com os colaboradores do projeto, de forma a levantar, analisar e selecionar todos os elementos que influenciam no desenvolvimento do projeto.
2	Seleção da Tecnologia a ser utilizada.	Estudos prospectivos das tecnologias existentes para a criação de bibliotecas digitais que mais atendem às necessidades.
3	Levantamento da documentação a ser disponibilizada na BDG para criação das coleções.	Como as coleções são definidas por tipologia dos documentos que comporão o acervo, foram levantadas amostras desses documentos para estabelecimento das coleções iniciais da biblioteca digital.
4	Análise das coleções para seleção dos metadados.	Com as coleções definidas, é preciso analisar os documentos para estabelecimento dos metadados para as coleções, e com isso facilitar a recuperação.
5	Customização da Tecnologia.	Com a tecnologia selecionada, e também as coleções e metadados selecionados, parte-se para a customização para implementar a biblioteca digital.
6	Adição de novas funcionalidades.	Ferramentas oferecem funcionalidades que podem agregar serviços aos sistemas de informação. Por isso pode-se adicionar novas funcionalidades à gestão de documentos digitais.
7	Implementação da identidade visual.	Sites de governo precisam adotar o modelo gov.br, motivo pelo qual é necessário colocar esta identidade visual na BDG.

Fonte: Elaboração própria.

Complementando as informações, conforme o modelo proposto por Mussi, Flores e Almeida (2021), o estudo foi realizado de julho a dezembro de 2024, sob a coordenação da

diretoria do Ibict, com a colaboração da Coordenação de Tecnologias para Informação (COTEC), em Brasília. Por ser a implantação de um sistema de informação, foi necessária a participação de bibliotecários, cientistas da informação, informáticos, entre outros, de forma a atender todas as necessidades, além de equipamentos básicos como computadores e servidores ligados à rede.

3 RESULTADOS E DISCUSSÕES

3.1 Seleção da Tecnologia

Uma das etapas essenciais na implementação do sistema é a escolha da tecnologia mais adequada. Diante disso, verificou-se a necessidade de prospectar tecnologias existentes para a criação de bibliotecas digitais, a fim de identificar a que melhor atendesse às demandas da BDG. Para isso, foi realizado um estudo para avaliar as características das tecnologias disponíveis e mapear os requisitos específicos do projeto.

As tecnologias analisadas foram o Dspace e o Omeka, tendo em vista a experiência da COTEC na implementação dessas tecnologias. A escolha justifica-se por suas similaridades: sistemas de código aberto, flexíveis e consistentes, organizados em coleções e compatíveis com padrões de metadados.

Após uma análise mais detalhada, o Omeka se mostrou a opção mais adequada, pois oferece uma interface mais intuitiva e acessível, facilitando a navegação e o uso da biblioteca. Além disso, o Omeka permite maior flexibilidade na personalização da plataforma, por meio de *plugins*, possibilitando a adaptação da estrutura conforme as demandas do projeto. Outro fator foi a leveza e menor complexidade técnica em comparação ao Dspace, tendo em vista que, apesar de ser uma plataforma robusta, requer mais recursos técnicos e operacionais. Já o Omeka, por ser mais simples e focado em coleções digitais, tem uma curva de aprendizado menos acentuada.

3.2 Definição das Coleções

A organização da BDG baseia-se na definição de coleções que agrupam documentos de acordo com a sua tipologia e função no contexto institucional. A estrutura teve como objetivo facilitar a recuperação da informação, garantindo acesso e navegabilidade ágeis.

As coleções foram definidas considerando a relevância de cada tipologia documental para o suporte à tomada de decisão, ao cumprimento das diretrizes e à transparência. Nesse sentido, os documentos foram categorizados conforme as características e finalidades específicas. O Quadro 2 apresenta essas tipologias e as suas definições.

Quadro 2 – Coleções

Nº	Tipologia	Descrição
1	Decreto	Norma expedida pelo Chefe do Poder Executivo, com força de lei, regulamentando políticas e diretrizes para a governança e gestão do Ministério da Ciência, Tecnologia e Inovação (MCTI) e do Ibict.
2	Plano	Instrumento de orientação das ações estratégicas, que estabelece compromissos para garantir o cumprimento dos objetivos do Instituto. Apresenta metas, objetivos, caminhos e propostas de governança.
3	Portaria	Ato administrativo do Diretor do Ibict, que disciplina questões operacionais e organizacionais no contexto do Instituto.
4	Regimento	Instituído por Portaria, é um conjunto de normas que disciplina o funcionamento da Instituição e das unidades administrativas, definindo a estrutura, competências e atribuições.
5	Outros	Inclui outros documentos que contribuem para a governança e gestão institucional do Ibict, como: editais, resoluções, memórias e manuais.

Fonte: Elaboração própria.

O desenvolvimento de coleções, por tipologia documental, encontra amparo em bibliotecas digitais, como apresentado por Basevi (2005) no modelo criado para o consórcio Biblioteca Digital Jurídica (BDJur). Tanto que grande parte das bibliotecas digitais do poder judiciário no Brasil segue esse modelo, no qual pode-se mesclar tipologia documental e formato documental.

3.3 Definição dos Metadados

A partir da amostra da documentação que compõe as coleções, extraiu-se os metadados necessários para descrevê-las e recuperá-las. Para isso, foi realizada uma análise das suas características, considerando as especificidades e a necessidade de padronização na representação da informação.

A escolha dos metadados buscou seguir o padrão Dublin Core, porém, devido à especificidade da documentação, foi necessário incluir elementos adicionais para garantir uma descrição mais precisa. Como mostra o Quadro 3, os metadados numerados de 1 a 10

correspondem à descrição do item em si, já os metadados numerados de 11 à 15 são dados da publicação, com o objetivo de tornar acessível na fonte original.

Quadro 3 – Metadados

Nº	Metadado	Decreto	Plano	Portaria	Regimento	Outras
1	Título	x	x	x	x	x
2	Autor	x	x	x	x	x
3	Vigência	x	x	x	x	x
4	Período		x		x	
5	Ementa	x		x	x	x
6	Data do ato	x		x	x	x
7	Número do processo	x		x	x	x
8	Assunto	x	x	x	x	x
9	Notas	x	x	x	x	x
10	Signatário	x		x	x	x
11	Instrumento de publicação	x		x	x	x
12	Local da publicação	x	x	x	x	x
13	Data da publicação	x	x	x	x	x
14	Ano da publicação	x	x	x	x	x
15	Texto completo da legislação	x		x	x	

Fonte: Elaboração própria.

Assim, além dos metadados básicos, como *Título*, *Autor*, *Assunto*, *Data do ato* e *Notas*, foram incorporados campos essenciais específicos dos atos normativos e administrativos, como *Vigência*, *Período*, *Ementa* e *Signatário*. Também foram adicionados metadados específicos da publicação, visto que alguns instrumentos podem ser considerados periódicos, como o Diário Oficial da União ou o Boletim de Serviços. Esses metadados incluem *Número do Processo*, *Instrumento de Publicação*, *Local da Publicação*, *Data da publicação* e *Ano*.

O uso do esquema de metadados Dublin Core (DC), principalmente o qualificado, encontra apoio em trabalhos como o de Macedo, Shintaku e de Brito (2015) e Andrade, Shintaku e Barros (2018), entre outros, no qual relatam a adequação desse padrão para

descrição de documentos governamentais. Essa opção pelo DC também dá-se pela possibilidade de interoperabilidade

3.4 Customização e Adição de Funcionalidades

Nesta etapa, a estrutura da plataforma começa a ser customizada, com a criação das coleções, dos tipos de itens e dos metadados. A partir disso, algumas modificações começam a ser necessárias para continuar a customização. Neste ponto, o Omeka se destaca com a adição de *plugins*, tanto que para Shintaku *et al.* (2018), esta adição é uma das grandes vantagens da utilização do Omeka. Gonçalves (2023) expõe que os *plugins* ampliam as funcionalidades, permitindo acrescentar de acordo com a demanda, além de ser simples a sua instalação.

Nesse sentido, foram instalados 23 *plugins* na BDG, cada um com funções específicas, voltadas para representação, disseminação, apresentação e busca dos itens na base de dados. Os *plugins* estão apresentados no Quadro 4.

Quadro 4 – Plugins

Nº	Plugin	Descrição
1	Bulk Metadata Editor	Permite editar vários registros ao mesmo tempo.
2	COinS	Adiciona metadados COinS nas páginas de itens, tornando-os legíveis pelo Zotero.
3	Collection Tree	Permite criar hierarquias entre as coleções, de modo a criar árvores de coleções.
4	CSV Export Format	Permite exportar registros no formato CSV.
5	CSV Import	Permite importar registros para o Omeka no formato CSV.
6	ElementManager	Possibilita gerenciar elementos de metadados após criados.
7	Element Types	Permite qualificar os metadados em tipos.
8	Exhibit Builder	Adiciona a possibilidade de criação de exposições.
9	Facets	Adiciona um bloco de navegação facetada nas páginas de busca dos itens.
10	Geolocation	Adiciona informação de geolocalização e mapas aos itens.
11	Heist	Facilita a responsividade entre desktop e dispositivos mobile.
12	Hide Elements	Ocultar elementos de metadados para despoluir a inserção de novos itens.

13	History Log	Cria um log de eventos, registrando todas as inserções, alterações e exclusões do sistema.
14	Item Order	Permite ordenar os itens dentro das coleções.
15	Item Relations	Permite adicionar relação entre os itens.
16	Locale Switcher	Fornece a opção de alternar a linguagem no sistema.
17	OAI-PMH Harvester	Coleta de metadados de provedores de dados OAI-PMH.
18	OAI-PMH Repository	Torna os itens da BDG exposta ao OAI-PMH.
19	Reports	Permite gerar relatórios em HTML e PDF.
20	Search By Metadata	Torna os valores dos metadados clicáveis para buscar elementos com a mesma descrição.
21	Simple Pages	Adiciona a opção de criação de páginas simples.
22	Simple Vocab Plus	Permite a criação de vocabulários controlados para cada campo de metadados.
23	Timeline	Adiciona a opção de criação de linhas do tempo.

Fonte: Elaboração própria.

Esses *plugins* foram selecionados de acordo com as necessidades identificadas, buscando eficiência na organização, na busca e na apresentação dos documentos. O conjunto de funcionalidades criou um ambiente flexível e consistente, alinhando práticas de gestão da informação com as necessidades da biblioteca.

3.4.1 Linha do Tempo

Uma funcionalidade que merece destaque é a inclusão da linha do tempo. No contexto da BDG, permite visualizar a evolução da legislação relacionada ao Ibict, oferecendo um panorama histórico. A linha do tempo da BDG começa em 1951, com o Registro das decisões da sexta sessão da Conferência Geral da Unesco, e está em constante atualização. A Figura 1 apresenta um slide da linha do tempo.

Figura 1 – Linha do tempo



Fonte: Elaboração própria .

A linha do tempo presente na BDG contribui na contextualização das normas que sustentam a atuação do Ibict ao longo dos anos. A sua implementação permite que os usuários identifiquem marcos históricos, explorando e compreendendo a história da instituição, além de fortalecer a transparência.

4 CONSIDERAÇÕES FINAIS

Estudos anteriores no Brasil sobre implementação de bibliotecas digitais têm uma longa história com o uso da ferramenta livre DSpace (Basevi, 2005; Ferreira, 2007; Medeiros; Dourado; Reis, 2013; entre outros). Tanto que Martins, Silva e Siqueira (2018) consideram o Dspace como uma boa opção para a criação desses sistemas de informação. Entretanto, o presente estudo optou pelo uso do Omeka, por se tratar de um Glam (Galleries, Libraries, Archives and Museums), na medida em que está destinado a atuar com a memória técnica, com documentação que é compartilhada pela biblioteca e arquivo.

Nesse caminho, o Omeka se apresenta adequado para gestão de coleções digitais de documentos de governança, mesmo sendo menos robusto que o DSpace, mas com maior flexibilidade e simplicidade. Diante disso, apresenta-se como uma boa opção para criação de bibliotecas digitais com menor complexidade e ação limitada, para disseminação de documentos em formato digital, utilizando as funcionalidades nativas do *software*.

REFERÊNCIAS

- ANDRADE, Morgana Carneiro de; SHINTAKU, Milton; BARROS, Patricia Pacheco de. Proposta de elementos de metadados para representação e recuperação de memória técnica: o caso da Rede Ufes-Rio Doce. **Cadernos de Biblioteconomia, Arquivística e Documentação**, Lisboa, n. 1, p. 41-58, 2018. DOI: 10.48798/cadernosbad.1918. Disponível em: <https://publicacoes.bad.pt/revistas/index.php/cadernos/article/view/1918>. Acesso em: 03 fev. 2025.
- BASEVI, Teresa. BDJur consortium-juridical digital library: implementing DSpace in the brazilian judiciary. In: CONFERENCE ON ELECTRONIC PUBLISHING, 2005, Leuven. **Proceedings [...]**. Leuven: Katholieke Universiteit Leuven, 2005. Disponível em: <https://elpub.architecturez.net/system/files/pdf/150elpub2005.content.pdf>. Acesso em: 02 fev. 2025.
- BRASIL. **Decreto Nº 9.203, de 22 de novembro de 2017**. Dispõe sobre a política de governança da administração pública federal direta, autárquica e fundacional. Brasília, DF: Presidência da República, 2017. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2017/decreto/d9203.htm. Acesso em: 03 fev. 2025.
- BRASIL. Presidência da República. Casa Civil. **Manual de redação da Presidência da República**. 3. ed. Brasília: Presidência da República, 2018. Disponível em: <https://www4.planalto.gov.br/centrodestudos/assuntos/manual-de-redacao-da-presidencia-da-republica/manual-de-redacao.pdf>. Acesso em: 15 mar. 2024.
- BRASIL. Tribunal de Contas da União. **Dez passos para a boa governança**. Brasília: TCU, Secretaria de Controle Externo da Administração do Estado, 2021. Disponível em: <https://portal.tcu.gov.br/publicacoes-institucionais/cartilha-manual-ou-tutorial/10-passos-para-a-bom-governanca>. Acesso em: 03 fev. 2025.
- FERREIRA, Sueli Mara Soares Pinto. Repositório institucional em comunicação: o projeto REPOSCOM implementado junto à Federação de Bibliotecas Digitais em ciências da comunicação. **Encontros Bibli**: revista eletrônica de biblioteconomia e ciência da informação, Florianópolis, v. 12, n. 1, p. 77-94, 2007. DOI: 10.5007/1518-2924.2007v12nesp1p77. Disponível em: <https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2007v12nesp1p77>. Acesso em: 03 fev. 2025.
- GONÇALVES, Maison. O Omeka do Tribunal de Justiça do Distrito Federal e dos Territórios. In: SHINTAKU, Milton; MORAES, Marcelo Hilario (org.). **Modernização Tecnológica do Ecosistema de Informação do Tribunal de Justiça do Distrito Federal e dos Territórios**. Brasília: Editora Ibict; TJDF, 2023. Cap. 7, p. 140-151. DOI: 10.22477/9786589167976.cap7. Disponível em: <https://labcotec.ibict.br/omp/index.php/edcotec/catalog/book/322>. Acesso em: 03 fev. 2025.

MACÊDO, Diego José; SHINTAKU, Milton; BRITO, Ronnie Fagundes de. Dublin core usage for describing documents in Brazilian government digital libraries. *In: INTERNATIONAL CONFERENCE ON DUBLIN CORE AND METADATA APPLICATIONS*, 2015, São Paulo.

Proceedings [...]. [S. l.]: Dublin Core Metadata Initiative, 2015. Disponível em: <https://dcpapers.dublincore.org/article/952137078>. Acesso em: 03 fev. 2025.

MARTINS, Dalton Lopes; SILVA, Marcel Ferrante; SIQUEIRA, Joyce. Comparação entre sistemas para criação de acervos digitais: análise dos softwares livres DSpace, EPrints, Fedora, Greenstone e Islandora a partir de novas dimensões analíticas. **InCID: Revista de Ciência da Informação e Documentação**, Ribeirão Preto, v. 9, n. 1, p. 52-71, 2018. DOI: 10.11606/issn.2178-2075.v9i1p52-71. Disponível em:

<https://www.revistas.usp.br/incid/article/view/134333>. Acesso em: 03 fev. 2025.

MEDEIROS, Ana Ligia Silva; DOURADO, Stella Moreira; REIS, Clea Mara Barradas. Biblioteca digital da Fundação Casa de Rui Barbosa: implantação do software DSpace. *In: CONGRESSO BRASILEIRO DE BIBLIOTECONOMIA, DOCUMENTAÇÃO E CIÊNCIA DA INFORMAÇÃO*, 25., 2013, Florianópolis. **Anais [...]**. São Paulo: Federação Brasileira de Associações de Bibliotecários, 2013, p. 1856-1871. Disponível em:

<http://repositorio.febab.org.br/items/show/2278>. Acesso em: 03 fev. 2025.

MONTEIRO, Silvana Drumond; CARELLI, Ana Esmeralda; PICKLER, Maria Elisa Valentin. A ciência da informação, memória e esquecimento. **DataGramaZero-Revista de Ciência da Informação**, [S. l.], v. 9, n. 6, p. 3-16, 2008. Disponível em: <https://brapci.inf.br/v/6382>. Acesso em: 03 fev. 2025.

MUSSI, Ricardo Franklin de Freitas; FLORES, Fábio Fernandes; ALMEIDA, Claudio Bispo de. Pressupostos para a elaboração de relato de experiência como conhecimento científico. **Revista praxis educacional**, Vitória da Conquista, v. 17, n. 48, p. 60-77, 2021. Disponível em:

http://educa.fcc.org.br/scielo.php?script=sci_arttext&pid=S2178-26792021000500060.

Acesso em: 03 fev. 2025.

PETERS, Brainard Guy. O que é governança? **Revista do TCU**, Brasília, ano 45, n. 127, maio/ago. 2013. Disponível em:

<https://revista.tcu.gov.br/ojs/index.php/RTCU/article/view/87>. Acesso em: 15 mar. 2024.

SHINTAKU, Milton et al. **Guia do usuário do Omeka**. Brasília: Ibict, 2018. Disponível em:

<https://repositorio.enap.gov.br/jspui/handle/1/3162>. Acesso em: 03 fev. 2025.



[4.0 Internacional](https://creativecommons.org/licenses/by/4.0/)

Esta licença permite compartilhamento, remixe, adaptação e criação a partir do trabalho, mesmo para fins comerciais, desde que sejam atribuídos créditos ao(s) autor(es). Conteúdos elaborados por terceiros, citados e referenciados nesta obra não são cobertos pela licença.

Como citar este trabalho:

BRAGA, Tiago Emmanuel Nunes; SHINTAKU, Milton; SANTOS, Cristiane Henrique; GONÇALVES, Maison Roberto Mendonça; COSTA, Mirele Carolina Souza Ferreira. Desenvolvimento de coleção digital voltada à governança em biblioteca digital. *In: WORKSHOP DE INFORMAÇÃO DADOS E TECNOLOGIA*, 8., 2025, Marília, SP. **Anais [...]**. Marília, SP: Universidade de Marília, 2025. DOI: <http://doi.org/10.22477/viii.widat.217>.



WIDaT 2025

VIII WORKSHOP DE INFORMAÇÃO,
DADOS E TECNOLOGIA

MARÍLIA • SP

O uso de *podcasts* por periódicos científicos na Ciência da Informação

The use of podcasts by scientific journals in Information Science

Diego André Andrade Neves¹



<http://lattes.cnpq.br/0489929643724568>



<https://orcid.org/0009-0003-4681-6750>

Milton Shintaku²



<http://lattes.cnpq.br/8605833104600600>



<https://orcid.org/0000-0002-6476-4953>

Resumo

A comunicação científica tem sido profundamente impactada pela digitalização e pela disseminação das mídias sociais. Os *podcasts* emergem como uma ferramenta relevante para divulgação científica, mas sua adoção por periódicos científicos na área de Ciência da Informação ainda não é explorada. Este estudo busca analisar como os periódicos dessa área utilizam os *podcasts*. O estudo levanta os *podcasts* com temas ligados à Ciência da Informação na busca dos vinculados à revista. Assim, pôde-se verificar que apenas a revista AtoZ, da Universidade Federal do Paraná (UFPR), publica um *podcast*, mas que muitas organizações fazem uso desse canal de forma oficial, o que pode apresentar potencialidade.

Palavras-chave: *podcasts*; comunicação científica; altmetria; Ciência Aberta; periódicos científicos.

Abstract

Scientific communication has been profoundly impacted by digitalization and the rise of social media. Podcasts have emerged as a relevant tool for scientific dissemination, but their adoption by scientific journals in the Information Science field remains unexplored. This study aims to analyze how journals in this field use podcasts. This research identifies podcasts related to Information Science to determine those linked to scientific journals. The findings reveal that only the journal AtoZ, from the Federal University of Paraná (UFPR), publishes a podcast. However, many organizations use this medium officially, indicating its potential for broader adoption.

Keywords: podcasts; scientific communication; altmetrics; Open Science; scientific journals.

¹ Discente mestrando pelo Programa de Pós-Graduação em Gestão da Informação (PPGGI) da Universidade Federal do Paraná (UFPR), Curitiba, PR, Brasil. Bolsista pelo Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), Brasília, DF, Brasil. diegoneves@ibict.br.

² Doutor em Ciência da Informação pela Universidade de Brasília (UnB), Brasília, DF, Brasil. Coordenador de Tecnologias para Informação (Cotec) no Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict). Docente Permanente do Programa de Pós-Graduação em Gestão da Informação (PPGGI) da Universidade Federal do Paraná (UFPR). shintaku@ibict.br.

1 INTRODUÇÃO

A rápida informatização impactou profundamente a comunicação científica, em especial os periódicos, que migraram do suporte físico para o digital com grande velocidade, mudando inclusive o modelo de negócio baseado na assinatura e distribuição física. Tanto que, o Brasil é um dos destaques na disseminação livre dos resultados de pesquisas (*Open Access*), sendo um dos pioneiros com o projeto *Scientific Electronic Library Online* (SciELO)³, lançada em 1998. Posteriormente, grande parte das revistas científicas nacionais passaram a publicar seus artigos gratuitamente em formato digital, utilizando o *software* livre *Open Journal Systems* (OJS), sendo o segundo país que mais utiliza a ferramenta.

Nesse novo cenário, as mídias sociais passaram a fazer parte da comunicação científica, criando assim uma nova métrica que estuda essa relação, a altmetria. Essa relação entre os periódicos e as mídias sociais tornou-se parte integrante. Araújo (2019), em editorial, comenta que estudiosos europeus discutem sobre os impactos na comunicação acadêmica e no comportamento dos pesquisadores. Nesse mesmo caminho, nota-se que cada vez mais a altmetria se apresenta fortalecida em um contexto da ciência aberta, integrando as mídias sociais como canais de divulgação científica.

Entretanto, há certos desafios que ainda não foram tratados, nem pela altmetria, a questão dos *podcasts*, ainda não contemplados pela métrica⁴, somente se o *podcast* for também publicado como *videocast* no YouTube é contabilizado. Com isso, cria-se o questionamento “*como os periódicos científicos da área da Ciência da Informação têm utilizado o formato podcast como estratégia de comunicação científica?*” na medida em que são significativamente utilizados na divulgação científica, como afirmam dos Santos e Barros (2023).

Assim, o estudo tem por objetivo verificar se as revistas científicas da Ciência da Informação brasileiras fazem uso desses canais para obter maior visibilidade e, com isso, discutir como essas mídias sociais se inserem na comunicação científica, vinculados aos

³ Disponível em: <https://www.scielo.org/pt/sobre-o-scielo/programa-scielo-modelo-scielo-de-publicacao-e-rede-scielo/>.

⁴ Disponível em: <https://www.altmetric.com/about-us/our-data/donut-and-altmetric-attention-score/>.

periódicos científicos. Apoiar o entendimento de como as revistas científicas estão se comportando no mundo digital e todas as suas possibilidades.

2 COMUNICAÇÃO CIENTÍFICA

A comunicação científica é um processo dinâmico que evoluiu ao longo dos anos, adaptando-se às novas tecnologias. É definida por Garvey (1979) como um conjunto de atividades associadas à produção, disseminação e uso da informação, a partir do momento em que o cientista concebe uma ideia para pesquisar, até que a informação sobre os resultados seja aceita como parte do conhecimento científico. Assim, os cientistas adotam tecnologias em suas atividades, incluindo na forma de disseminação.

Targino (2000) aponta que a comunicação científica é indispensável à atividade científica, pois:

“[...] permite somar os esforços individuais dos membros das comunidades científicas. Eles trocam continuamente informações com seus pares, emitindo-as para seus sucessores e/ou adquirindo-as de seus predecessores. É a comunicação científica que favorece ao produto (produção científica) e aos produtores (pesquisadores) a necessária visibilidade e possível credibilidade no meio social em que produto e produtores se inserem.” (Targino, 2000, p. 10).

Para Bueno (2010), a comunicação científica e a divulgação científica se viabilizam a partir de canais ou veículos que também exibem características distintivas bem definidas. Aponta-se isso pois, ao adentrar na camada digital, a divulgação científica se desmembra em distintos canais e formatos.

2.1 A Comunicação Científica na Era Digital

Nesse caminho, o formato digital impactou profundamente na divulgação e disseminação da informação, podendo, com isso, comparar o surgimento dessa modalidade à invenção dos tipos móveis por Gutemberg. Abriu-se novas possibilidades de apresentar a informação, passando do textual para o multimídia. No campo da comunicação científica, as revistas passaram a ser digitais, com acesso via *web*, mas mantendo o formato tradicional textual.

No cenário da internet, com grande disputa informacional, a comunicação científica tem feito uso das mídias sociais. Isso fez com que fosse criado uma métrica especializada para verificar o seu impacto neste ambiente, a altmetria. Como visto por Piccoli e Stecanela (2023), a altmetria volta-se a medir o impacto social da informação científica, por meio da mensuração desta informação nas mídias sociais. Entretanto, as mídias sociais apresentam uma grande variedade de tipos, dentre tem-se os *podcasts*.

2.2 Os Podcasts

O *podcast*, conforme definição de Costa e Silva (2024), consiste em um formato de mídia de áudio digital que se realiza por meio da distribuição desses arquivos pela internet, usando o recurso de disparo de seu *Really Simple Syndication* (RSS) pelos seus agregadores. Nos últimos anos, este formato evoluiu para o audiovisual, também conhecido por *videocast*. Os *podcasts* têm se destacado na divulgação da Ciência, pois permitem transmissão dinâmica de conteúdo e acessibilidade ampliada. A popularidade dos *podcasts* na ciência cresce exponencialmente. Costa e Silva (2024) ainda apontam que, internacionalmente e nacionalmente, os *podcasts* estão sendo objeto de pesquisa e atenção por parte da academia.

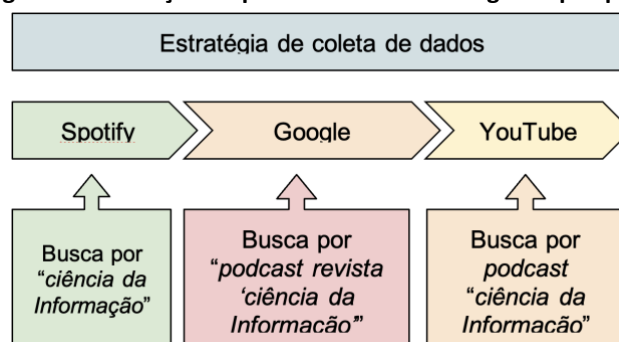
3 METODOLOGIA

Este estudo adota uma abordagem qualitativa, quantitativa e descritiva, abrangendo levantamento de periódicos científicos da Ciência da Informação indexados em bases de dados e análise da existência de *podcasts* vinculados a essas revistas. Devido ao objetivo de verificar se as revistas científicas da Ciência da Informação brasileiras fazem uso dos *podcasts* a estratégia de coleta de dados segue o seguinte percurso (Figura 1).

Como o *Spotify* é a rede social que abriga somente os *podcasts*, buscou-se apenas pelo tópico “ciência da informação”. Como muitos *podcasts* também usam a sua versão *videocast*, buscou-se por “*podcast*” e “*ciência da informação*”, de forma a pegar todos os vídeos com o tema. Por fim, a busca no google, mais abrangente, restringiu a busca para

três argumentos: “*podcast revista ‘ciência da informação’*”, tendo sido utilizado a categoria de pesquisa “*Todas*” e o filtro “*Pesquisar páginas em português*”, com maior possibilidade de recuperação de dados.

Figura 1 - Ilustração do percurso da metodologia da pesquisa



Fonte: Elaboração própria.

4 RESULTADOS

A primeira busca, obteve como resultado diversos *podcasts* científicos, porém, sem um foco no contexto pesquisado. Esses resultados confirmam os estudos de dos Santos e Barros (2023) e Figueira e Bevilacqua (2022), que indicam que os *podcasts* são utilizados para divulgação científica. Ao clicar na opção “*Mostrar tudo*”, expandiu-se os resultados, porém, ainda mantendo uma diversidade de ocorrências generalizadas dentro da ciência como um todo, sem nenhum foco na área da Ciência da Informação, exceto um único resultado que foi o *podcast* “*Introdução a Ciência da Informação*”.

Ao dar prosseguimento na pesquisa, foi utilizada a plataforma Google. Como resultado, houveram novamente ocorrências diversas, embora desta vez dentro do foco da Ciência da Informação, listando aproximadamente 51.500 resultados. O que diversificou foram os formatos de resultados, consistindo entre artigos científicos, matérias de sites e blogs, mas sem a consistência dos *podcasts* propriamente ditos.

Souza e Santos (2018) apontam que os meios de comunicação são fundamentais para que informações sejam consumidas, enfatizando ainda que os avanços tecnológicos possibilitaram a comunicação digital. Então, como terceira tentativa, utilizou-se da categoria de pesquisa “*Vídeos*”, mostrando bastante resultados da proposta, no caso

diversos episódios do *podcast* da *Revista AtoZ: novas práticas em informação e conhecimento*⁵, da Universidade Federal do Paraná (UFPR).

Araújo *et al.* (2023) argumentam que:

O aperfeiçoamento das Tecnologias Digitais de Informação e Comunicação (TDIC) têm impactado e ampliado continuamente o processo de divulgação científica, por meio do uso de redes sociais online como Instagram, Facebook e Twitter, dentre outros canais e formatos, a exemplo dos serviços de streaming que apresentam a ferramenta de *podcast* (Araújo *et al.*, 2023, p. 2).

Então, seguindo este fluxo de pesquisa, mais uma tentativa foi feita, porém na plataforma YouTube. Foram utilizados os mesmos termos no buscados da plataforma de vídeos, e os resultados consistiram em ocorrências de episódios de *podcasts* da referida revista, *AtoZ*. Rolando mais a tela, outros vídeos inconsistentes e diversos foram surgindo, perdendo-se mais uma vez o foco da proposta desta pesquisa.

Assim, ao voltar na plataforma Google e verificando novamente mais resultados além no rolar da tela, conseguiu-se encontrar uma ocorrência que muito chamou atenção. Era uma matéria do site do Sistema de Bibliotecas e Informação (SiBi) da Universidade Federal do Rio de Janeiro (UFRJ), com o nome de “*Podcasts sobre Biblioteconomia e Ciência da Informação*”.

O mesmo, embora datado em 10 de maio de 2021, levantava uma relação de todos os *podcasts* de Ciência da Informação (e sua área correlata, Biblioteconomia), conforme mostra a Figura 2.

Na matéria foram listados todos os possíveis *podcasts* existentes no Brasil que abordam a Ciência da Informação, embora nem todos provenientes de periódicos. Complementando a lista apresentada, pode-se verificar quais os *podcasts* vinculados a revistas, como apresentado no Quadro 1.

⁵ Disponível em: <https://revistas.ufpr.br/atoz>.

Figura 2 – Matéria sobre *podcasts* de Ciência da Informação no SiBi/UFRJ

Fonte: Captura de tela da matéria no site do Sibi/UFRJ (2025)⁶.

Quadro 1 - Relação dos respectivos *podcasts* da lista com suas vinculações institucionais e se são oriundos de revistas científicas de Ciência da Informação

Nº	Nome do Podcast	Vinculação	Revista?
1	Além das Estantes	Não informado	Não
2	Biblio do Calmon	Pessoal	Não
3	Biblioteconomia em Foco	FACC - UFRJ	Não
4	Biblioteconomia para Concursos	Pessoal	Não
5	Biblioteco	Não informado	Não
6	Biblioteca em Prosa	Não informado	Não
7	BIBLIOQUÊ	Universidade Federal do Amazonas (UFAM)	Não
8	BiblioCast	Não informado	Não
9	Biblio S.A	PET Biblioteconomia UNIRIO	Não
10	Biblio Tubers	Não informado	Não
11	Cibercultura De Pierre Lévy	Universidade Estadual da Paraíba (UEPB)	Não

⁶ Disponível em: <https://www.sibi.ufrj.br/index.php/inicio/471-podcasts-sobre-biblioteconomia-e-ciencia-da-informacao>. Acesso em: 21 fev. 2025.

12	Cimplifica - O primeiro podcast brasileiro de Ciência da Informação	Não informado	Não
13	Competência em Informação Instrumental	Faculdade de Informação e Comunicação (FIC) da Universidade Federal do Amazonas (UFAM)	Não
14	Consciência e Informação	Programa de Pós-Graduação em Ciência da Informação (PPGCI) da Universidade Federal da Paraíba (UFPB)	Não
15	DescompliCast	Departamento de Ciência da Informação da Universidade Federal da Paraíba (UFPB)	Não
16	Foca na Info	Não informado	Não
17	Nepomuceno Estratégia Digital Nepô	Não informado	Não
18	OldSchool Biblio	Não informado	Não
19	Panema Cultural	Cursos de Biblioteconomia e Documentação e Arquivologia da Universidade Federal Fluminense	Não
20	Planejamento em Unidades de informação	Não informado	Não
21	Podcast InformarÉ	Conselho Regional de Biblioteconomia da 8ª Região (CRB-8) - SP	
22	PodCI	Pessoal	Não
23	Podcast do LTI Digital	Universidade Federal da Bahia (UFBA)	Não
24	Revista AtoZ	Programa de Pós-Graduação em Gestão da Informação (PPGGI) da Universidade Federal do Paraná (UFPR)	Sim
25	Transitando na Biblio	Conselho Regional de Biblioteconomia do Estado do Rio de Janeiro (CRB-7 RJ)	Não
26	Transitando na Biblio Podcast	Conselho Regional de Biblioteconomia do Estado do Rio de Janeiro (CRB-7 RJ)	Não

Fonte: Elaboração própria.

Mesmo que a presença de *podcasts* de revistas ainda seja tímida (uma única revista), há uma grande quantidade vinculado a organizações e instituições, dando oficialidade a estes canais. Assim, pode-se verificar que, sendo um canal utilizado para divulgação científica, ainda é pouco utilizado como canal de disseminação, com vinculação à revista, podendo apresentar oportunidades para outros estudos.

4.1 Os Podcasts da Revista AtoZ e a Inteligência Artificial (IA)

Visto que essa única revista é a Revista AtoZ, faremos então um apanhado descritivo de sua atuação dentro do contexto da comunicação científica, como seu formato, periodicidade, objetivos e episódios do *podcast* desta revista.

O *podcast* da Revista AtoZ foi lançado em julho de 2021, e é uma ação do projeto de extensão “*Ciência Aberta e a Gestão da Informação Científica*” da Universidade Federal do Paraná (UFPR). Sua periodicidade é quinzenal e seu formato é audiovisual (*videocast*). Os encontros são compostos por 1 (um) entrevistador ou entrevistadora (ou mais de 1, conforme dinâmica acordada entre a equipe de produção). O objetivo do mesmo consiste em abordar tudo que está relacionado à ciência aberta, à gestão da informação e à pesquisa científica, especialmente no que tange a tecnologia informacional para tanto a disseminação quanto a atuação do corpo docente, discente e pesquisador dentro do cenário dessa gestão. Os convidados são desde professores e discentes (inclusive de pós graduação) à atuantes externos ao escopo da UFPR.

No que se trata do uso da Inteligência Artificial (IA), o *podcast* usa desta tecnologia para a disseminação de seus episódios. Além dos mesmos serem postados pelo Youtube, são também publicados pela plataforma *beacon.ai*, um sistema que ajuda criadores de conteúdo a gerir a sua presença online e engajar o seu público.

A plataforma *beacon.ai* utiliza a IA para compilar links diversos em um único link. No caso do *podcast* da Revista AtoZ, este *link* divulga os episódios pelas plataformas: *Spotify*, *Deezer*, *Apple Podcasts*, *Amazon Music*, *Audible*, *Castbox*, *Pocket Cast*, *Podcast Addict*, *PodVine*, *RadioPublic*, e *TuneIn*.

5 CONSIDERAÇÕES FINAIS

Os resultados obtidos evidenciam que o uso de *podcasts* na comunicação científica ainda é pouco explorado no âmbito dos periódicos científicos da Ciência da Informação no Brasil. A ausência de mensuração direta pela altmetria, salvo quando o conteúdo é disponibilizado como *videocast* no YouTube, limita o reconhecimento e a visibilidade desse formato como ferramenta de divulgação acadêmica. Apesar dessa

limitação, a pesquisa identificou iniciativas pontuais de uso de *podcasts* por instituições acadêmicas, embora, no caso, apenas 1 vinculada a revista científica, destacando a Revista AtoZ da UFPR.

Ademais, a análise dos resultados obtidos a partir das buscas nas plataformas *Spotify*, *Google* e *YouTube* demonstra que há um crescente interesse pelo formato, mas sem uma estruturação consolidada nos periódicos científicos. Diante disso, é essencial que mais estudos sejam conduzidos para avaliar o potencial dos *podcasts* como meio de disseminação da ciência. Além disso, recomenda-se que as métricas altmétricas no que tange a mídias sociais *versus* comunicação científica sejam aprimoradas para incorporar esse formato de forma mais abrangente. É possível que, com o avanço das tecnologias digitais e a crescente aceitação dos *podcasts* no meio acadêmico, essa ferramenta venha a desempenhar um papel cada vez mais relevante na comunicação científica.

REFERÊNCIAS

- ARAÚJO, Ronaldo Ferreira de. O impacto das mídias sociais para revistas científicas da área da saúde. **Acta Paulista de Enfermagem**, São Paulo, v. 32, n. 1, 2019. DOI: <https://doi.org/10.1590/1982-0194201900001>. Disponível em: <https://www.scielo.br/j/ape/a/jzKLTX7htBnQTJ9FwD9HNDb>. Acesso em: 21 fev. 2025.
- ARAÚJO, Joana Ferreira de; SILVA, Alzira Karla Araújo; AUTRAN, Marynice de Medeiros Matos; TELMO, Flávia de Araújo. Divulgação científica e podcast: disseminação do conhecimento científico na Ciência da Informação. **Brazilian Journal of Information Science: research trends**, [S. l.], n. 17, p. 1-23. 2023. DOI: <https://doi.org/10.36311/1981-1640.2023.v17.e023046>. Disponível em: <https://dialnet.unirioja.es/servlet/articulo?codigo=9161388>. Acesso em: 28 fev. 2025.
- BUENO, Wilson Costa. Comunicação científica e divulgação científica: aproximações e rupturas conceituais. **Informação & Informação**, Londrina, v. 15, n. 1, p. 1-12, 2010. DOI: <https://doi.org/10.5433/1981-8920.2010v15n1espp1>. Disponível em: <https://ojs.uel.br/revistas/uel/index.php/informacao/article/view/6585>. Acesso em: 09 abr. 2025.
- COSTA, Camila Martineli; SILVA, Fabiano Couto Corrêa da. Podcasts como ferramenta para comunicação científica: um estudo sobre a divulgação da Ciência da Informação. **AtoZ: novas práticas em informação e conhecimento**, Curitiba, v. 13, p. 1-13, 2024. DOI: <https://doi.org/10.5380/atoz.v13i0.89539>. Disponível em: <https://revistas.ufpr.br/atoz/article/view/89539>. Acesso em: 21 fev. 2025.

FIGUEIRA, Ana Cristina Peixoto; BEVILAQUA, Diego Vaz. Podcasts de divulgação científica: levantamento exploratório dos formatos de programas brasileiros. **Revista Eletrônica de Comunicação, Informação & Inovação em Saúde**, [S. l.], v. 16, n. 1, 2022. DOI: <https://doi.org/10.29397/reciis.v16i1.2427>. Disponível em: <https://www.reciis.icict.fiocruz.br/index.php/reciis/article/view/2427>. Acesso em: 28 fev. 2025.

GARVEY, William D. **Communication, the essence of science**: facilitating information exchange among librarians, scientists, engineers, and students. New York: Pergamon Press, 1979.

MUELLER, Suzana Pinheiro Machado. Popularização do conhecimento científico. **DataGramZero**: Revista de Ciência da Informação, [S. l.], v. 3, n. 2, 2002. Disponível em: <https://brapci.inf.br/v/5354>. Acesso em: 21 fev. 2025.

SANTOS-D'AMORIM, Karen. A comunicação científica em movimento: das origens aos debates atuais. **Brazilian Journal of Information Science**, Marília, v. 15, p. e02103, 2021. DOI: <https://doi.org/10.36311/1981-1640.2021.v15.e02103>. Disponível em: <https://revistas.marilia.unesp.br/index.php/bjis/article/view/11468>. Acesso em: 21 fev. 2025.

PICCOLI, Marcia Speguen de Quadros; STECANELA, Nilda. Popularização da ciência: uma revisão sistemática de literatura. **Educação e Pesquisa**, São Paulo, v. 49, p. e253818, 2023. Disponível em: <https://www.scielo.br/j/ep/a/55yQ3zb8pLrwPD3kcdyQFdk>. Acesso em: 21 fev. 2025.

SANTOS, Sara Pires; BARROS, Adriano David Monteiro. Podcast como instrumento de divulgação científica: uma análise bibliométrica. **Estudos em Comunicação**, [S. l.], n. 36, 2023. DOI: <https://doi.org/10.25768/1646-4974n36a09>. Disponível em: <https://ojs.labcom-ifp.ubi.pt/ec/article/view/1157>. Acesso em: 21 fev. 2025.

SOUZA, Vanessa Lima de; SANTOS, Jacques Fernandes. Comunicação midiática: influência, consumo e moda nas redes sociais. **RIOS: Revista Científica da Faculdade Sete de Setembro**, [S. l.], v. 12 n. 17. 2018. Disponível em: <https://www.publicacoes.unirios.edu.br/index.php/revistarios/article/view/341>. Acesso em 28 fev. 2025.

TARGINO, Maria das Graças. Comunicação científica: uma revisão de seus elementos básicos. **Informação & Sociedade**, João Pessoa, v. 10, n. 2, 2000. Disponível em: <https://periodicos.ufpb.br/ojs/index.php/ies/article/view/326>. Acesso em: 09 abr. 2025.



[4.0 Internacional](https://creativecommons.org/licenses/by/4.0/)

Esta licença permite compartilhamento, remixe, adaptação e criação a partir do trabalho, mesmo para fins comerciais, desde que sejam atribuídos créditos ao(s) autor(es). Conteúdos elaborados por terceiros, citados e referenciados nesta obra não são cobertos pela licença.

Como citar este trabalho:

NEVES, Diego André Andrade; SHINTAKU, Milton. O uso de podcasts por periódicos científicos na Ciência da Informação. *In*: WORKSHOP DE INFORMAÇÃO DADOS E TECNOLOGIA, 8., 2025, Marília, SP. **Anais [...]**. Marília, SP: Universidade de Marília, 2025. DOI: <http://doi.org/10.22477/viii.widat.232>.

Uso de IA para criação de Ontologias: uma proposição para o domínio “Saúde Indígena”

Use of AI for Ontology creation: a proposition for the "Indigenous Health" domain

Luiz Rogério Lopes Silva¹

 <http://lattes.cnpq.br/6377367856927679>

 <https://orcid.org/0000-0002-7457-9778>

Marcelo Schiessi²

 <https://lattes.cnpq.br/4402964117780625>

 <https://orcid.org/0000-0002-3895-5844>

Milton Shintaku³

 <http://lattes.cnpq.br/8605833104600600>

 <https://orcid.org/0000-0002-6476-4953>

Resumo

Introdução: Propõe o uso de Inteligência Artificial (IA), especificamente a LLM ChatGPT-4, para a criação de Ontologia de domínio em “saúde indígena” de forma semi-automatizada, a fim de auxiliar o desenvolvimento de um sistema de informação para a Secretaria de Saúde Indígena (SESAI). **Metodologia:** Adotou-se o framework LLMs4OL, adaptando-o para o contexto específico. Inicialmente, foi definido o escopo do domínio e coletado um corpus controlado de documentos referentes à SESAI. Em seguida, foram elaborados *prompts* para o ChatGPT-4, com o objetivo de extrair conceitos e relações do corpus. As respostas do modelo foram avaliadas e os *prompts* foram refinados iterativamente para melhorar a precisão e a completude da ontologia. **Resultados:** Uma série de *prompts* são disponibilizados para criação de ontologias em outros domínios. A ontologia criada é composta por 180 axiomas, dos quais 54 são axiomas de declaração e 58 são axiomas lógicos. Há 44 classes distintas, 11 propriedades de objeto e 2 propriedades de anotação, além de 14 indivíduos identificados. **Conclusão:** O uso da IA demonstrou eficácia na extração de termos, relações e na criação de uma estrutura ontológica inicial, mostrando um potencial promissor para criação de ontologias. Entretanto, a supervisão humana garantiu a adequação cultural e a precisão dos dados, provando que a combinação de IA e expertise humana é uma alternativa para organização do conhecimento.

Palavras-chave: inteligência artificial; ontologia; chatGPT-4; saúde indígena; aprendizagem de ontologia.

¹ Doutor em Gestão da Informação, Universidade Federal do Paraná (UFPR), Curitiba, PR, Brasil. Professo colaborador do Programa de Pós-graduação em Gestão da Informação da Universidade Federal do Paraná (UFPR), Curitiba, PR, Brasil. Pesquisador (Bolsista), Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), Brasília, DF, Brasil. luiz.lopes@ufpr.br.

² Doutor em Ciência da Informação, Universidade de Brasília (UnB), Brasília, DF, Brasil. Pesquisador (Bolsista), Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), Brasília, DF, Brasil. joseschiessi@ibict.br.

³ Doutor em Ciência da Informação, Universidade de Brasília (UnB), Brasília, DF, Brasil. Coordenador de Tecnologias para Informação, Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), Brasília, DF, Brasil. shintaku@ibict.br.

Abstract

Introduction: This paper proposes the use of Artificial Intelligence (AI), specifically the LLM ChatGPT-4, for the semi-automated creation of a domain ontology in "indigenous health," in order to assist in the development of an information system for the Special Secretariat of Indigenous Health (SESAI). **Methodology:** The LLMs4OL framework was adopted, adapting it to the specific context. Initially, the domain scope was defined, and a controlled corpus of documents related to SESAI was collected. Then, prompts were designed for ChatGPT-4 to extract concepts and relations from the corpus. The model's responses were evaluated, and the prompts were iteratively refined to improve the accuracy and completeness of the ontology. **Results:** A series of prompts is provided for creating ontologies in other domains. The created ontology consists of 180 axioms, 54 of which are declaration axioms and 58 are logical axioms. There are 44 distinct classes, 11 object properties, and 2 annotation properties, as well as 14 identified individuals. **Conclusion:** The use of AI proved effective in extracting terms, relations, and creating an initial ontological structure, demonstrating promising potential for ontology creation. However, human supervision ensured cultural appropriateness and data accuracy, proving that the combination of AI and human expertise is a viable alternative for knowledge organization.

Keywords: artificial intelligence; ontology; chatGPT-4; indigenous health; ontology learning

1 INTRODUÇÃO

A Secretaria Especial de Saúde Indígena (SESAI) é responsável por coordenar e executar a Política Nacional de Atenção à Saúde dos Povos Indígenas no Brasil, o que inclui a gestão de obras e edificações destinadas ao atendimento médico nas aldeias. Nesse sentido, Benevides, Portillho e Nascimento (2014) afirmam que a SESAI é responsável por planejar, executar e avaliar ações relacionadas ao saneamento e edificações nos Distritos Sanitários Especiais Indígenas (DSEI).

Araújo e Felipe (2021) apresentam, mesmo que de forma discreta, a relação entre as obras de saneamento e melhoria da saúde indígena nas aldeias, assim como as diferenças entre os distritos indígenas, distribuídos por todo o território nacional. Com isso, ressaltam a necessidade de um sistema de informação que gerencie as atividades relacionadas às edificações, que podem estar relacionadas ao saneamento ou à saúde.

Diante deste problema, a SESAI firmou projeto de pesquisa com o Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict) para propor um ecossistema de informação que possa melhorar a eficiência das suas ações relacionadas a obras de engenharia nas aldeias. Uma das etapas do projeto é criar uma ontologia de domínio da saúde indígena, processo visto como facilitador para estruturar um entendimento formal e compartilhado do campo.

Todavia, criar uma ontologia exige esforço em analisar e extrair conhecimento especializado do domínio. A SESAI, suas coordenações e os Distritos Sanitários Especiais Indígenas (DSEIs) produzem volume extenso de documentos e informações relacionados

à gestão, assistência e monitoramento da saúde indígena, incluindo os processos de obras de engenharia, o que por sua vez, intensifica a complexidade em organizar o conhecimento.

No contexto deste artigo, os autores encontram nos estudos de *Ontology Learning*, ou aprendizado de ontologias, uma alternativa para reduzir a dependência de trabalho manual intensivo e de expertise específica para desenvolver uma ontologia do domínio "Saúde Indígena". A literatura aponta que uma das maneiras de automatizar a extração e estruturação do conhecimento contido em textos variados é o uso de Modelos de Linguagem de Grande Escala (LLMs). Tais modelos são capazes de identificar padrões e extrair informações relevantes mesmo de textos não estruturados, organizando conceitos, relações e regras que do modo tradicional demandaria mais tempo e dedicação.

Assim, o objetivo deste trabalho é apresentar os procedimentos usados na criação de uma ontologia de domínio da saúde indígena que utilizou de LLM para automatizar a extração e estruturação de conhecimento. O trabalho se justifica tendo em vista a necessidade de equilibrar a precisão e a riqueza cultural exigidas pela saúde indígena com a eficiência operacional necessária para a gestão em larga escala. A aplicação de inteligência artificial em contextos socioculturais complexos, como é o caso da SESAI, promove uma inovação que pode servir de modelo para outras iniciativas da organização da informação e do conhecimento.

2 ONTOLOGY LEARNING

Aprendizado de Ontologias (AO) é uma disciplina que combina Inteligência Artificial e engenharia de conhecimento para construir automaticamente uma ontologia a partir de fontes textuais estruturadas ou semiestruturadas (Giglou; D'Souza; Auer, 2023). As tarefas do AO incluem identificar conceitos, significados e relações em textos, além de extrair axiomas, o que pode reduzir a necessidade de construção manual (Giglou; D'Souza; Auer, 2023).

Na literatura, modelos distintos têm sido propostos para auxiliar na engenharia de ontologias. *Ontology Learning Layer Cake* é um modelo introduzido por Philipp Cimiano, em 2006, que estrutura o processo de AO em camadas sequenciais. Começa com

a extração de termos lexicais e prossegue com a identificação de sinônimos, construção de taxonomias, extração de relações não hierárquicas e, finalmente, formulação de axiomas e regras. Cada camada do modelo concentra-se em um aspecto diferente da ontologia, desde a identificação básica de termos até a definição de relações complexas e restrições. Esse modelo permite uma análise detalhada e modular, sendo útil para compreender as etapas individuais do processo de criação de ontologias.

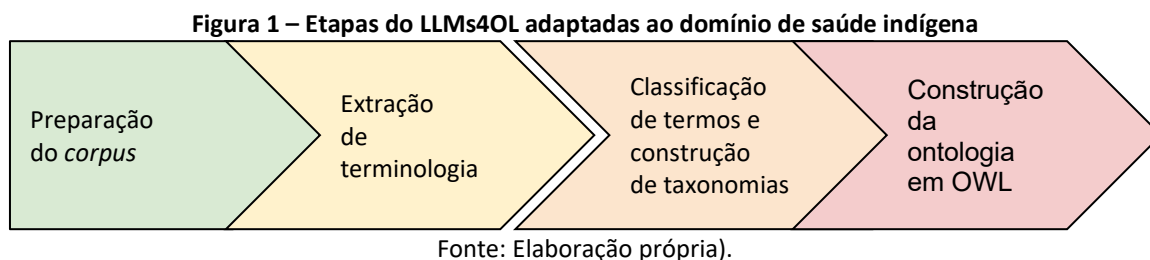
O modelo *End-to-End Ontology Learning com Large Language Models* (LLMs), proposto por Lo *et al.* (2024), busca automatizar todas as etapas do AO de forma contínua. Utiliza técnicas de Processamento de Linguagem Natural (PLN) para extrair e formalizar conhecimento diretamente das fontes de dados, sem a necessidade de intervenção manual em cada etapa. Este modelo se diferencia por tentar capturar todas as nuances necessárias para a construção de ontologias em uma única fase, o que pode acelerar significativamente o processo de desenvolvimento.

O paradigma *Large Language Models for Ontology Learning* (LLMs4OL), desenvolvido por Giglou, D'Souza e Auer, em 2023, utiliza LLMs para realizar tarefas específicas de AO. Isso inclui preparação de corpus, extração de terminologia, classificação de termos e construção de taxonomias. O LLMs4OL reconhece que, embora os LLMs possam realizar essas tarefas, a validação por especialistas é essencial para garantir a relevância e a precisão dos axiomas gerados.

Para o desenvolvimento da ontologia de gestão de obras e edificações em comunidades indígenas, o paradigma LLMs4OL foi escolhido como o modelo a ser seguido. Essa escolha se justifica pela capacidade dos LLMs de processar dados textuais e pela importância da validação especializada, oferecendo um equilíbrio entre automação e expertise humana.

3 PROCEDIMENTOS METODOLÓGICOS

Conforme o objetivo proposto, este trabalho é uma pesquisa aplicada adaptando modelos da IA para a construção de uma ontologia específica. Assim, o estudo segue o fluxo apresentado na Figura 1, com base no paradigma LLMs4OL.



A primeira etapa, a preparação do corpus, envolveu a coleta de documentações oficiais da SESAI relacionadas a edificações, que abrangem informações sobre Estrutura e Organização Institucional, Fluxos Processuais, Políticas Públicas e Sistemas de Informação. Esses documentos foram integrados em um único conjunto de dados por meio de um *script* em *Python*, permitindo que o método fosse replicável em diferentes contextos.

Com o conjunto de dados preparado, foram realizados testes utilizando quatro modelos de LLMs: *ChatGPT-4*, *Maritalk*, *QWEN* e *Deepseek*. Assim, para avaliar o desempenho desses modelos no processo de criação de ontologias, adotamos um conjunto de métricas que envolvem tanto aspectos de qualidade semântica quanto de eficiência computacional. Entre as principais métricas utilizadas, destacam-se:

- Cobertura conceitual: verifica se os termos, classes e relações extraídos cobrem os conceitos principais de “saúde indígena”;
- coerência e consistência dos axiomas: verifica se há conflitos ou duplicações nos axiomas para garantir a viabilidade da ontologia;
- eficiência de processamento: inclui tanto o tempo de resposta dos modelos na extração de conceitos e relações quanto a quantidade de passos de refinamento de prompts necessários para atingir um resultado satisfatório.

Nas análises empíricas, o ChatGPT-4 superou os demais em termos de eficácia – melhor equilíbrio entre precisão e cobertura conceitual – e eficiência – menor tempo médio de iterações até a geração de um resultado coerente. Além disso, foi o modelo que apresentou menor incidência de relações redundantes ou inconsistentes, reduzindo o esforço de validação manual.

A definição das categorias conceituais relevantes para o domínio de “saúde indígena” foi guiada por um processo semiestruturado que combinou:

- a) Análise prévia de documentos e políticas da SESAI: estrutura organizacional, fluxos processuais e termos técnicos recorrentes;
- b) metodologias formais de *Ontology Learning*, em especial o paradigma LLMs4OL, que orienta a elaboração de ontologias a partir de corpus;
- c) validação por especialistas: discussões com pesquisadores do tema “saúde indígena” para revisar e confirmar a pertinência das categorias.

A técnica utilizada para a criação dos *prompts* uniu referências formais, como as etapas sugeridas pelo LLMs4OL, com ajustes ad hoc baseados na resposta inicial do modelo. Em outras palavras, partimos de uma estrutura metodológica clara – organizando a coleta de dados, extração de termos e modelagem – e, a cada iteração, refinamos os comandos de entrada – *prompts* – conforme analisávamos a qualidade das saídas geradas. Esse ciclo de realimentação contínua buscou melhorar a aderência dos resultados à realidade do domínio.

Portanto, na etapa de extração de terminologia, os termos relevantes foram identificados a partir do corpus preparado. O *ChatGPT-4* foi empregado para auxiliar na extração desses termos, aproveitando sua capacidade de processar e entender a linguagem natural. Em seguida, na classificação de termos, os conceitos extraídos foram agrupados em categorias conceituais relevantes, identificando classes ou tipos a que os termos pertencem.

A construção de taxonomias foi realizada ao estabelecer relações hierárquicas “is-a” entre os conceitos identificados. Além disso, foram extraídas relações semânticas não-taxonômicas, capturando relações complexas entre os conceitos.

As relações semânticas não taxonômicas foram identificadas durante o processo de extração e classificação de termos, tendo como base:

- a) Verbos e expressões verbais como executar, gerir, fiscalizar que indicam ações ou vínculos funcionais entre os conceitos. Por exemplo: “Distrito Sanitário Especial Indígena (DSEI)” –[executa]→ “Obras de Saneamento”. Isso significa que o DSEI realiza, na prática, a execução ou coordenação de obras de saneamento no território indígena;
- b) expressões funcionais ou de dependência como por exemplo, depende de, pertence a, está contido em ou oferece suporte a. Desse modo, a relação

“Equipe de Saúde” – [pertence a] → “DSEI” pode ser inferida da frase: “a equipe de saúde está formalmente vinculada ao DSEI”.

Quando identificadas relações mais complexas (p. ex., que envolvem três ou mais conceitos), utilizamos regras como “Se (X *fazParteDe* Y) e (Y *supervisionadoPor* Z) então (X *supervisionadoPor* Z)”, formalizando-as em formato de *Semantic Web Rule Language* (SWRL) ou em axiomas OWL adicionais. Tal regra representa a herança da relação de supervisão, típica do contexto de unidades de saúde subordinadas a uma secretaria maior.

Por fim, foram criados axiomas e regras para definir restrições que regem a ontologia, garantindo sua consistência e validade. O conhecimento foi representado formalmente em triplas *Resource Description Framework* (RDF) e *Web Ontology Language* (OWL), permitindo uma estrutura que pode ser facilmente compartilhada e reutilizada. O detalhamento de cada uma das etapas é apresentado a seguir.

4 PROPOSIÇÃO DE ONTOLOGIA COM ChatGPT-4

O presente estudo é uma proposição e, por isso, a formação do *corpus* deu-se com um pequeno conjunto de documentos, de forma a facilitar os processos seguintes, na medida em que há conhecimento sobre eles. Como primeiro resultado, o Quadro 1 apresenta os *prompts* utilizados em cada etapa e os resultados alcançados.

Quadro 1 – Etapas de desenvolvimento

Etapa	Prompts	Resultado Alcançado	Exemplo
Extração de Termos	Analise o texto fornecido e extraia a maior quantidade possível de entidades nomeadas (NER), incluindo pessoas, localidades, organizações, funções e atividades. Liste os principais substantivos, verbos (convertidos para o infinitivo) e siglas e acrônimos (proponha a forma expandida, se disponível ou inferível). Organize cada lista em ordem alfabética, numerando os itens e evitando duplicatas. Adicione a quantidade total de termos do resultado.	Extraíram-se 282 termos dos dados, categorizados em 95 substantivos, 70 verbos e 117 siglas.	O termo “DSEI Yanomami” foi classificado como organização/localidade; verbo “monitorar” indica ação típica das equipes de saúde; e sigla “SIS (Sistema de Informação em Saúde)”, interpretada como sistema computacional.

Geração de Taxonomia	Com base nos termos extraídos no Passo 1, classifique cada termo em tipos conceituais apropriados (ex: 'instituição', 'procedimento', 'equipamento', 'pessoa', 'organização', 'conceito abstrato'). Crie uma hierarquia taxonômica, definindo relações de superclasse-subclasse (ex: 'organização' > 'empresa' > 'startup'). Se um termo for polissêmico, crie ramificações adicionais. Estructure o resultado em tabela com: 1. Termo original. 2. Tipo conceitual associado. 3. Relação hierárquica (supertipo e subtipo). Evite repetições, exceto para termos com múltiplos significados. Adicione a quantidade total de termos classificados do resultado.	Criação de uma estrutura taxonômica de 31 termos.	Foi criada a relação "Organização > Secretaria > SESAI", pois a SESAI foi reconhecida como um tipo específico de Secretaria na hierarquia.
Extração de Relações Não Taxonômicas	Análise todos os termos e a taxonomia gerada nos Passos 1-2. Identifique relações semânticas não taxonômicas entre os tipos conceituais. Organize as relações em uma tabela no formato: #(número da relação)– Tipo A – [relação] – Tipo B (com breve justificativa ou evidência textual). Evite duplicações e inclua apenas relações semanticamente válidas. Adicione a quantidade total de relações encontradas.	Identificação de 54 propriedades relacionais.	Identificou-se, por exemplo, 'DSEI' –[implementa]→ 'Programa de Saneamento', mostrando uma vinculação operacional; e 'Obra de Edificação' –[exige]→ 'Licença Ambiental', indicando a necessidade legal.
Extração de Axiomas	Com base nos passos anteriores, defina todos os axiomas possíveis. Crie regras (ex: SWRL) para comportamentos complexos, como: <i>Se X trabalha em Y, e Y pertence a Z, então X está relacionado a Z</i> Organize as definições numeradas em tabelas: 1. Axiomas de Classe. 2. Axiomas de Propriedade. 3. Regras Unárias/Binárias. 4. Regras N-árias. Garanta consistência lógica e evite contradições. Adicione a quantidade total de axiomas encontrados.	Identificação de 13 axiomas.	Axioma "(DSEI SubClassOf Organização)" formaliza que cada DSEI é uma instância de "Organização".
Criação de triplas RDF	Com base em todos os termos, relações e axiomas dos Passos 1-4,	Identificação de 36 triplas.	<DSEI_Yanomami> <rdfs:subClassOf> <DSEI> -

	gere triplas RDF no formato <sujeito> <predicado> <objeto> Evite duplicações e organize as triplas em seções ou listas, agrupando por tipo de relação ou termo. Se houver ambiguidade, indique-a claramente. Adicione a quantidade total de triplas geradas.		<SESAI> <supervisiona> <DSEI> - <ObraEdificacao> <exige> <LicencaAmbiental> Cada tripla espelha uma relação fundamental no domínio, seja ela hierárquica (subclasse) ou funcional (coordena, supervisiona).
Criação da ontologia em owl	Com base nas triplas RDF do Passo 5, organize as classes em uma estrutura hierárquica (superclasses e subclasses). Defina atributos (propriedades) relevantes para cada classe e crie axiomas para restringir e descrever melhor as classes e propriedades. Formule regras para comportamentos complexos ou inferências que não possam ser expressas apenas com axiomas OWL. Garanta que as definições sejam coerentes e não gerem contradições lógicas e gere a ontologia OWL mais completa possível, com base nos passos até aqui. Além disso, estenda a ontologia com instâncias (indivíduos) adicionais e inclua as triplas RDF listadas anteriormente.	Ontologia gerada em OWL.	Na ontologia final, por exemplo, "SESAI" e "DSEI" são classes com restrições de domínio e alcance, além de incluir indivíduos como <DSEI_Xavante> e <DSEI_Yanomami>. As propriedades de objeto "coordena", "supervisiona" e "exige" possuem domínios e alcances compatíveis, prevenindo inconsistências lógicas.
Exportação e visualização da ontologia	Considere a ontologia construída nos Passos 1-6, contendo classes, propriedades, axiomas e regras. Salve a ontologia em um arquivo OWL (ex: 'minha_ontologia.owl'). Gere, também, um código de visualização para Graphviz, criando um diagrama simplificado com: 1. Nós representando classes. 2. Arestas representando relações (taxonômicas e não taxonômicas). 3. Cores e layouts intuitivos para facilitar a interpretação.	Arquivo OWL gerado e código para GraphViz fornecido.	O código para Graphviz cria nós para cada classe principal (por exemplo, "SESAI", "DSEI", "ObraSaneamento", "LicencaAmbiental") e arestas rotuladas ("coordena", "supervisiona"), evidenciando a hierarquia de classes e as conexões centrais do domínio de "saúde indígena".

Fonte: Elaboração própria.

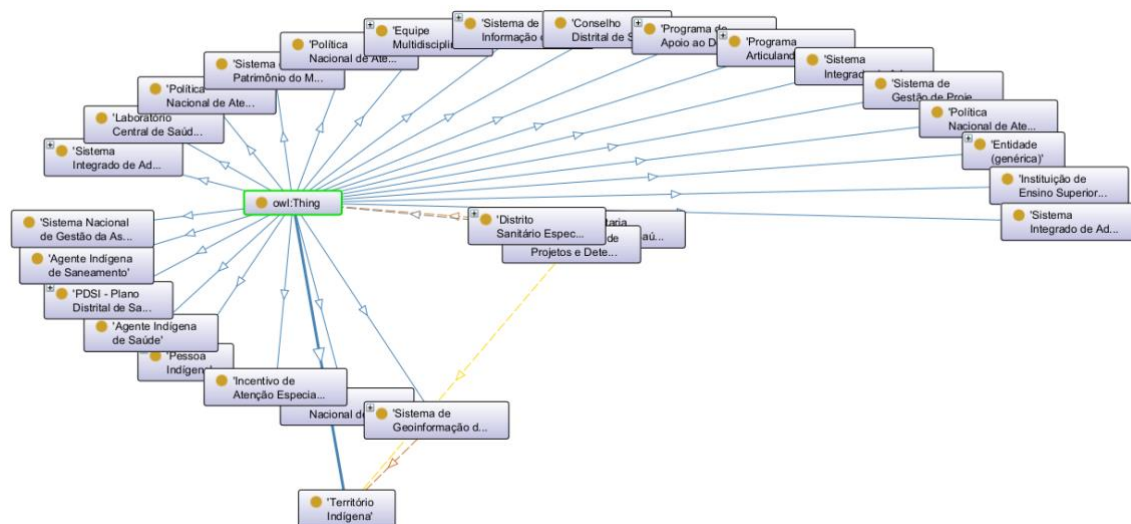
Inicialmente, a capacidade de compreensão e representação lógico-semântica do ChatGPT-4 evidenciou um resultado parcial e satisfatório do domínio. A ferramenta demonstrou eficácia na definição de termos e na criação da ontologia, analisando

documentos rapidamente e superando métodos manuais. Sua escalabilidade permitiu a extração de informações, identificando padrões e relações complexas.

Em um segundo momento, o achado de pesquisa esteve no contexto da engenharia de *prompts* e na necessidade de refinar continuamente as instruções fornecidas ao ChatGPT-4. Os autores avaliaram que a simples aplicação dos princípios LLMs4OL foi insuficiente para o domínio "Saúde Indígena", exigindo que fosse otimizada a interação para melhorar a qualidade das respostas.

No que compete à ontologia gerada em OWL, Figura 2, o artefato é composto por 180 axiomas, dos quais 54 são axiomas de declaração e 58 são axiomas lógicos. Há 44 classes distintas, 11 propriedades de objeto e 2 propriedades de anotação, além de 14 indivíduos identificados. A ontologia também detalha os tipos de axiomas presentes, como *SubClassOf* e *ClassAssertion*. No entanto, foi identificada uma diferença entre o número de termos, relações e axiomas gerados anteriormente pelo ChatGPT-4 e o que a ferramenta exportou em OWL. Neste momento, os pesquisadores não investigaram o motivo da discrepância, mas consideram realizar outros testes na sequência para tentar identificar a lógica utilizada.

Figura 2 – Representação da ontologia



Fonte: Elaboração própria.

No campo ético da pesquisa, questões de transparência e soberania do conhecimento merecem atenção. A natureza proprietária dos algoritmos e a falta de divulgação dos dados de treinamento da LLM criam uma "caixa preta", dificultando a

análise crítica de vieses e pressupostos na ontologia gerada. Isso compromete a confiança na representatividade fiel do conhecimento do domínio. Para tanto, o próprio modelo LLMs4OL exige a validação de especialistas, o que aconteceu graças ao conhecimento dos próprios pesquisadores sobre o domínio. Vale reforçar que a equipe de desenvolvedores está envolvida com a SESAI desde o ano de 2024, participando de eventos, analisando documentos e visitas *in loco* aos distritos indígenas.

5 CONSIDERAÇÕES FINAIS

A criação de uma ontologia de saúde indígena para a SESAI, utilizando o ChatGPT, identificou avanços e preocupações. A capacidade da LLM de processar volumes de dados, identificar relações e gerar representações de conhecimento em tempo reduzido é um avanço. A ferramenta demonstrou eficácia na extração de termos, relações e criação de uma estrutura ontológica inicial. Por outro lado, a questão da explicabilidade é um dilema: a opacidade no processo de geração das respostas, principalmente na criação do arquivo em OWL, provocou inquietações e receios.

Inicialmente, observou-se uma discrepância entre a quantidade de classes, axiomas e relações gerados pelos *prompts* nas fases anteriores e os gerados no OWL. Tal situação exige investigações futuras para evitar que a criação de ontologias por meio de LLMs seja uma “caixa preta”. Dada a novidade do tema, ainda há uma lacuna na realização de testes que observem o desempenho das ferramentas na proposição de terminologias confiáveis para domínios específicos. Um gerador pode criar um texto que parece plausível, mas isso não garante que ele esteja utilizando conhecimento sobre o domínio, o que pode ser resultado do pouco rigor na seleção e confiabilidade das fontes mapeadas. Não basta reduzir o tempo e o esforço humano envolvidos no processo, é necessário um processo de geração de conhecimento que seja suficientemente transparente e confiável.

Após a criação da ontologia, a avaliação focou em critérios como precisão, pertinência, abrangência, correção e a estruturação da informação. A versão inicial do artefato demonstrou ser apropriada para representar o domínio em questão. Contudo, os testes prosseguem com o objetivo de aprimorar essa primeira versão e assegurar que as representações sejam confiáveis e verdadeiras para o domínio da saúde indígena. A

equipe responsável valoriza e busca manter um equilíbrio entre a automação proporcionada pela IA e a supervisão e co-criação humanas.

REFERÊNCIAS

ARAÚJO, A. N. S.M.; FELIPE, E. C. Análise qualitativa das ações de saneamento e saúde em áreas indígenas. **Brazilian Journal of Development**, Curitiba, v. 7, n. 5, p. 53203-53213, 2021. DOI: <http://doi.dx/10.34117/bjdv.v7i5.30478>. Disponível em: <https://ojs.brazilianjournals.com.br/ojs/index.php/BRJD/article/view/30478>. Acesso em: 10 mar. 2025.

BENEVIDES, L.; PORTILLO, J. A. C.; NASCIMENTO, W. F. A atenção à saúde dos povos indígenas do Brasil: das missões ao subsistema. **Tempus Actas de Saúde Coletiva**, Brasília, v. 8, n. 1, p. 29-39, mar. 2014. DOI: <http://doi.dx/10.18569/tempus.v8i1.1450>. Disponível em: <http://www.tempusactas.unb.br/index.php/tempus/article/view/1450>. Acesso em: 10 mar. 2025.

CIMIANO, P. **Ontology learning and population from text: algorithms, evaluation and applications**. New York: Springer, 2006. DOI: <http://doi.dx/10.1007/978-0-387-39252-3>. Disponível em: <http://link.springer.com/10.1007/978-0-387-39252-3>. Acesso em: 10 mar. 2025.

GIGLOU, H. B.; D'SOUZA, J.; AUER, S. LLMs4OL: large language models for ontology learning. In: PAYNE, T. R.; PRESUTTI, V.; QI, G.; VILLALÓN-POVEDA, M.; STOILLOS, G.; HOLLINK L.; KAOUDI, Z.; CHENG, G.; LI, J. (ed). **The semantic web: ISWC 2023**. Athens: Springer, 2023. v. 14265, p. 408-427. DOI: http://doi.dx/10.1007/978-3-031-47240-4_22. Disponível em: https://link.springer.com/10.1007/978-3-031-47240-4_22. Acesso em: 10 mar. 2025.

LO, A.; JIANG, A. Q.; LI, W.; JAMNIK, M. End-to-end ontology learning with large language models. In: CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS, 38., 2024, Vancouver. **Conference [...]**. Vancouver: NeurIPS, 2024. DOI: <http://doi.dx/10.48550/arXiv.2410.23584>. Disponível em: <http://arxiv.org/abs/2410.23584>. Acesso em: 10 mar. 2025.



[4.0 Internacional](https://creativecommons.org/licenses/by/4.0/)

Esta licença permite compartilhamento, remixe, adaptação e criação a partir do trabalho, mesmo para fins comerciais, desde que sejam atribuídos créditos ao(s) autor(es). Conteúdos elaborados por terceiros, citados e referenciados nesta obra não são cobertos pela licença.

Como citar este trabalho:

SILVA, Luiz Rogério Lopes; SCHIESSL, Marcelo; SHINTAKU, Milton. Uso de IA para criação de Ontologias: uma proposição para o domínio "Saúde Indígena". In: WORKSHOP DE INFORMAÇÃO DADOS E TECNOLOGIA, 8., 2025, Marília, SP. **Anais [...]**. Marília, SP: Universidade de Marília, 2025. DOI: <https://doi.org/10.22477/viii.widat.260>.



WIDaT 2025

VIII WORKSHOP DE INFORMAÇÃO,
DADOS E TECNOLOGIA

MARÍLIA • SP

Aprimoramento do Annif para indexação automática de publicações em português

Enhancement of Annif for automatic indexing of publications in portuguese

Bruno Carlos da Cunha Costa¹



<http://lattes.cnpq.br/8162384236816488>



<https://orcid.org/0000-0001-9655-8698>

Ana Carolina Simionato Arakaki²



<http://lattes.cnpq.br/9896600626524397>



<https://orcid.org/0000-0002-0140-9110>

Renan Luiz da Silva Nascimento³



<http://lattes.cnpq.br/9394350252301981>



<https://orcid.org/0009-0001-9785-8833>

Resumo

A indexação automática desempenha um papel essencial na organização e recuperação da informação, especialmente em bibliotecas e repositórios institucionais que lidam com grandes volumes de documentos. O presente estudo investiga a aplicação de uma ferramenta baseada em inteligência artificial para a indexação automática de publicações em língua portuguesa, com o objetivo de aprimorar a precisão e a coerência dos descritores atribuídos aos documentos. Para isso, foram conduzidos testes com uma ferramenta desenvolvida originalmente para a indexação de textos em outro idioma, avaliando sua capacidade de adaptação a um novo contexto linguístico. A metodologia adotada incluiu a realização de testes exploratórios iniciais, seguidos da criação de um conjunto de treinamento específico composto por trinta mil documentos extraídos de bases institucionais. O treinamento foi conduzido de maneira incremental, permitindo a avaliação da evolução do desempenho da ferramenta. Os resultados indicaram que a qualidade das sugestões geradas pela ferramenta está diretamente relacionada à diversidade e estrutura do conjunto de treinamento. Inicialmente, a indexação apresentou baixa precisão para documentos em língua portuguesa, mas, após o refinamento dos dados e a ampliação do conjunto de treinamento, observou-se uma melhora significativa na coerência dos descritores atribuídos. O estudo demonstra que a adaptação da ferramenta para a indexação de documentos em língua portuguesa é viável, desde que sejam realizados ajustes metodológicos no treinamento do modelo. Conclui-se que o aprimoramento contínuo do conjunto de treinamento pode fortalecer a padronização da indexação,

¹ Doutor, Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro, RJ, Brasil Professor, Instituto Federal de Educação, Ciência e Tecnologia (IFRJ), Rio de Janeiro, RJ, Brasil. brunocosta@ibict.br.

² Doutora, Universidade Estadual Paulista Júlio de Mesquita Filho, UNESP, Brasil. Coordenadora de Serviços Bibliográficos, Instituto Brasileiro de Informação em Ciência e Tecnologia (COBIB/IBICT), Brasília, DF, Brasil. anaarakaki@ibict.br.

³ Graduação em andamento em Estatística, Universidade de Brasília, UnB, Brasília, DF, Brasil. Bolsista, Instituto Brasileiro de Informação em Ciência e Tecnologia (COBIB/IBICT), Brasília, DF, Brasil. renannascimento@ibict.br.

melhorar a recuperação da informação e reduzir a necessidade de intervenção manual em bibliotecas e repositórios institucionais.

Palavras-chave: indexação automática; Inteligência Artificial; processamento de linguagem natural.

Abstract

Automatic indexing plays a fundamental role in information organization and retrieval, especially in libraries and institutional repositories that manage large volumes of documents. This study investigates the application of an artificial intelligence-based tool for the automatic indexing of publications in Portuguese, aiming to improve the accuracy and coherence of descriptors assigned to documents. To achieve this, tests were conducted with a tool originally developed for indexing texts in another language, assessing its ability to adapt to a new linguistic context. The methodology included initial exploratory tests, followed by the creation of a specific training set consisting of thirty thousand documents extracted from institutional databases. Training was conducted incrementally, allowing for the evaluation of the tool's performance over time. The results indicated that the quality of the generated suggestions is directly related to the diversity and structure of the training set. Initially, the indexing showed low accuracy for documents in Portuguese. However, after refining the data and expanding the training set, a significant improvement was observed in the coherence of the assigned descriptors. The study demonstrates that adapting the tool for indexing documents in Portuguese is feasible, provided that methodological adjustments are made in model training. It is concluded that the continuous improvement of the training set can enhance indexing standardization, improve information retrieval, and reduce the need for manual intervention in libraries and institutional repositories.

Keywords: automatic indexing; Artificial Intelligence; natural language processing.

1 INTRODUÇÃO

O *Annif*⁴ foi desenvolvido em 2017 como um protótipo experimental para automatizar a indexação documental, utilizando *scripts* em *Python* e um índice do *Elasticsearch* para associar termos extraídos dos títulos de documentos a conceitos de um vocabulário controlado. Diferentemente dos sistemas tradicionais de recuperação da informação, em que o usuário busca documentos a partir de um tema, o *Annif* inverte essa lógica: recebe um documento como entrada e retorna os assuntos mais relevantes atribuídos ao seu conteúdo.

Seu nome, um anagrama de "*Finna*", serviço de descoberta usado em bibliotecas finlandesas, reflete essa inovação. Em 2018, a ferramenta foi ampliada com a adoção dos *frameworks Flask* e *Connexion*, permitindo a criação de uma interface *Web* interativa e uma API REST, o que facilitou sua integração a outros sistemas. A classificação automática do *Annif* baseia-se na execução de diferentes *backends*, ou seja, algoritmos que operam

⁴ Disponível em: <https://annif.org/> e seu código <https://github.com/NatLibFi/Annif>.

isoladamente ou combinados para melhorar a precisão das sugestões. Essa arquitetura modular permite sua adaptação a diversos vocabulários controlados e conjuntos documentais (Suominen, 2019).

Sua principal vantagem é a flexibilidade, possibilitando treinamento com diferentes bases de conhecimento e tornando-se útil para bibliotecas, repositórios institucionais e outros sistemas que necessitam de uma indexação escalável e padronizada. No Brasil, o Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict) tem explorado o *Annif* para aprimorar bases como a Rede Bibliodata e o Catálogo Coletivo Nacional de Publicações Periódicas (CCN), buscando maior interoperabilidade e normalização de registros bibliográficos por meio da Inteligência Artificial e do aprendizado de máquina.

A eficiência do *Annif* depende de conjuntos de treinamento adequados à área do conhecimento em questão. Embora permita testes com documentos de exemplo, seu desempenho é otimizado com dados específicos. O formato recomendado de entrada é .txt, e a escolha do *backend* influencia diretamente a qualidade da indexação. Para este estudo, foi adotado o *Term Frequency-Inverse Document Frequency* (TF-IDF), um método estatístico amplamente utilizado para atribuir pesos a termos, identificando os mais representativos em um conjunto documental.

A introdução do *Annif* na indexação e catalogação documental representa um avanço para a Biblioteconomia, ao automatizar a atribuição de descritores e reduzir a subjetividade do processo manual, frequentemente influenciado por variáveis cognitivas e interpretativas dos indexadores (Fujita, 1988). Como a indexação é essencial para a organização da informação e afeta diretamente a recuperação de documentos, a consistência e precisão dos termos atribuídos tornam-se aspectos críticos (Rubi e Fujita, 2003). Ferramentas automatizadas, como o *Annif*, mitigam essas limitações, proporcionando maior uniformidade terminológica e escalabilidade na indexação documental.

A qualidade da recuperação da informação depende da coerência dos metadados, sendo que sua estruturação adequada melhora a busca e a interoperabilidade entre bases de dados (Haynes, 2018). Conforme Pomerantz (2015), a

integração de modelos estatísticos e vocabulários controlados permite que sistemas como o Annif gerem sugestões mais precisas e alinhadas ao conteúdo dos documentos. Além disso, a normalização terminológica, conforme Svenonius (2000), é um fator decisivo para evitar ambiguidades e variações linguísticas, garantindo um controle mais eficaz dos termos indexados.

A automação da indexação não apenas reduz o esforço manual dos profissionais da informação, mas também possibilita a análise de grandes volumes documentais com maior rapidez e precisão. A adoção de ferramentas baseadas em inteligência artificial fortalece a infraestrutura da gestão da informação, tornando os processos de organização e recuperação da informação mais eficientes e beneficiando pesquisadores, bibliotecários e usuários finais.

Dessa forma, este estudo tem como objetivo avaliar o desempenho do *Annif* na indexação automática de publicações em português e propor estratégias para aprimorar sua eficácia, por meio do treinamento com dados especializados extraídos de bases de informação do Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict).

2 PROCEDIMENTOS METODOLÓGICOS

Este estudo adota uma abordagem experimental e aplicada, com o objetivo de avaliar a eficácia do *Annif* na indexação automática de publicações em português e propor estratégias para aprimorar seu desempenho. A pesquisa segue um delineamento quantitativo e qualitativo, combinando testes controlados da ferramenta com análise crítica dos resultados. A metodologia foi estruturada em três etapas principais: preparação do corpus de treinamento, configuração da indexação e análise dos resultados gerados.

A investigação caracteriza-se como um estudo experimental, no qual diferentes configurações do *Annif* foram testadas para determinar sua capacidade de indexação em português. Além disso, trata-se de uma pesquisa aplicada, pois visa desenvolver uma solução prática para aprimorar a indexação automatizada em bibliotecas e repositórios nacionais.

Inicialmente, foram conduzidos testes exploratórios com publicações de exemplo fornecidas pelo próprio *Annif*, todas em inglês. O objetivo era validar o funcionamento da ferramenta em um idioma previamente treinado e verificar a coerência das sugestões geradas. Os resultados foram satisfatórios, com alta correspondência entre os termos sugeridos e os conteúdos dos documentos.

Na segunda fase, o *Annif* foi testado com publicações científicas em português, selecionadas a partir das revistas do Ibict. No entanto, os resultados foram significativamente inferiores, com baixa precisão na atribuição dos descritores. Esse desempenho insatisfatório sugeriu dificuldades na adaptação do modelo para a língua portuguesa, possivelmente devido à ausência de um vocabulário treinado para esse idioma.

Diante disso, foi elaborado um corpus de treinamento especializado, contendo aproximadamente 30 mil publicações em português extraídas das bases do Ibict. Os documentos foram estruturados no formato *tab-separated values* (.tsv,) exigido pelo *Annif*, no qual cada linha contém o título e, quando disponível, o resumo da publicação, associados a *Uniform Resource Identifiers* (URIs) correspondentes aos descritores indexados na ontologia da ferramenta.

Para otimizar a indexação automática, foi realizada uma análise dos vocabulários utilizados pelo Ibict, mapeando-os para as ontologias do *Annif* a fim de garantir maior coerência na atribuição dos termos. Esse processo permitiu que o treinamento fosse ajustado às particularidades semânticas do português e às necessidades específicas das bibliotecas e repositórios nacionais.

O *backend* selecionado para a etapa de treinamento foi o *Term Frequency-Inverse Document Frequency* (TF-IDF), amplamente utilizado para análise de relevância de termos em documentos. Esse método estatístico atribui pesos às palavras com base na frequência com que ocorrem dentro do corpus, permitindo que o *Annif* identifique os termos mais representativos para cada publicação.

O treinamento do modelo foi realizado de forma incremental, no período de julho de 2024 a março de 2025, com avaliações semanais à medida que novos dados eram incorporados ao corpus, cabe destacar que o modelo enfrentou desafios específicos

relacionados à adaptação ao idioma português, o que exigiu estratégias adicionais de refinamento. Durante o processo de treinamento incremental, foram aplicadas técnicas de limpeza e normalização textual, padronização de grafias e tratamento de siglas recorrentes nas publicações acadêmicas. Além disso, buscou-se equilibrar o corpus entre diferentes áreas temáticas para evitar o viés de indexação em domínios mais representados. Também foram realizadas análises de erros a partir das sugestões inadequadas geradas, o que possibilitou ajustes iterativos nos parâmetros do *backend* TF-IDF e na estrutura dos arquivos de entrada. Essas ações contribuíram significativamente para aprimorar a sensibilidade do modelo às especificidades linguísticas do português e aumentar a relevância dos descritores sugeridos.

A avaliação do desempenho do *Annif* foi baseada em testes comparativos, conduzidos semanalmente com um conjunto fixo de cinco publicações em português. Para cada teste, as sugestões geradas foram analisadas em relação à coerência semântica, precisão e recall dos termos atribuídos. Os principais indicadores utilizados na análise foram: Precisão, a partir do percentual de descritores sugeridos que realmente correspondem aos conteúdos dos documentos, revocação: percentual de termos relevantes efetivamente recuperados pelo modelo, e F1-score: métrica que combina precisão e revocação, avaliando o equilíbrio entre ambas.

Além das métricas quantitativas, foi realizada uma avaliação qualitativa, na qual especialistas em Biblioteconomia e Ciência da Informação examinaram a pertinência dos termos atribuídos pelo *Annif*, considerando sua adequação às práticas de indexação bibliográfica.

3 RESULTADOS

A saída do *Annif* segue um formato simples e estruturado, retornando, por padrão, dez sugestões de assuntos organizadas em ordem decrescente de relevância. A pontuação atribuída a cada termo varia de 0 a 1, indicando o grau de confiança do sistema na associação entre o assunto sugerido e o conteúdo analisado. Valores mais próximos de 1 sinalizam maior precisão na indexação, permitindo que os usuários avaliem a qualidade das recomendações.

Para o treinamento do modelo, foram utilizadas três fontes de dados: Bibliodata, publicações do Encontro Nacional de Pesquisa em Ciência da Informação (ENANCIB) – Grupo de Trabalho 8 (Dado, Informação e Tecnologia) e uma seleção de textos da Sociedade Brasileira de Computação (SBC). Essa diversidade garantiu um conjunto representativo de documentos, cobrindo distintas áreas do conhecimento e fortalecendo a base de aprendizado da ferramenta. No entanto, nas primeiras semanas de testes, a escassez de dados corretamente formatados comprometeu os resultados, evidenciando a necessidade de um corpus mais robusto.

As limitações iniciais tornaram-se evidentes na análise das sugestões geradas. Em diversos casos, os termos atribuídos não apresentavam correspondência coerente com o conteúdo dos textos, especialmente em documentos que abordavam conceitos técnicos e específicos. Essa inconsistência revelou dificuldades na adaptação do modelo ao contexto semântico dos documentos, reforçando a importância de um treinamento mais refinado. A seguir, na Figura 1, são apresentadas imagens ilustrando essa limitação, com títulos das publicações, palavras-chave e sugestões geradas pelo *Annif*.

Figura 1: Sugestões do Annif para três publicações utilizando o “arquivo de treinamento inicial”

Título: Publicação de dados de pesquisa no contexto da comunicação científica: cenário terminológico-conceitual		Título: Arquitetura da informação para publicação de dados abertos conectados	
Palavras-chave: dados de pesquisa; ciência aberta; comunicação científica		Palavras-chave: metadados de negócio, dados abertos conectados, arquitetura d. informação para processos de negócio, web semântica	
Annif:		Annif:	
<http://www.yso.fi/onto/ysop3056>	databases 0.3221	<http://www.yso.fi/onto/ysop3056>	databases 0.3193
<http://www.yso.fi/onto/ysop3286>	social work 0.1598	<http://www.yso.fi/onto/ysop16380>	music therapy 0.1959
<http://www.yso.fi/onto/ysop4>	fiction and poetry 0.1292	<http://www.yso.fi/onto/ysop5643>	neuroses 0.1818
<http://www.yso.fi/onto/ysop2945>	learning 0.1156	<http://www.yso.fi/onto/ysop4>	fiction and poetry 0.1740
<http://www.yso.fi/onto/ysop13084>	nature 0.1153	<http://www.yso.fi/onto/ysop14783>	national literature 0.1682
<http://www.yso.fi/onto/ysop27522>	fanzines 0.1087	<http://www.yso.fi/onto/ysop1217>	travel 0.1682
<http://www.yso.fi/onto/ysop3333>	alternative movements 0.1087	<http://www.yso.fi/onto/ysop27522>	fanzines 0.1594
<http://www.yso.fi/onto/ysop608>	racism 0.1087	<http://www.yso.fi/onto/ysop3333>	alternative movements 0.1594
<http://www.yso.fi/onto/ysop14524>	algorithms 0.1057	<http://www.yso.fi/onto/ysop608>	racism 0.1594
<http://www.yso.fi/onto/ysop183>	research 0.1041	<http://www.yso.fi/onto/ysop6702>	credits 0.1534

Título: Aplicando o ChatGPT na condução de revisões sistemáticas da literatura	
Palavras-chave: revisão sistemática da literatura; ChatGPT; inteligência artificial	
Annif:	
<http://www.yso.fi/onto/ysop27522>	fanzines 0.2967
<http://www.yso.fi/onto/ysop3333>	alternative movements 0.2967
<http://www.yso.fi/onto/ysop608>	racism 0.2967
<http://www.yso.fi/onto/ysop4>	fiction and poetry 0.2391
<http://www.yso.fi/onto/ysop1178>	entrepreneurs 0.2295
<http://www.yso.fi/onto/ysop10315>	scientific journals 0.2234
<http://www.yso.fi/onto/ysop3286>	social work 0.2138
<http://www.yso.fi/onto/ysop2921>	religions 0.1981
<http://www.yso.fi/onto/ysop8113>	literature 0.1964
<http://www.yso.fi/onto/ysop21530>	software development 0.1886

Fonte: Elaboração própria.

Inicialmente, apenas a sugestão *databases* apresentou alguma relação com a publicação original, enquanto os demais termos atribuídos não possuíam conexão

4 CONSIDERAÇÕES FINAIS

Os testes demonstraram que a qualidade e diversidade do arquivo de treinamento inicial e final, são fatores determinantes para a precisão das sugestões geradas pelo *Annif*. Inicialmente, os resultados iniciais para publicações em português foram insatisfatórios, devido à escassez de dados adequados e à necessidade de ajustes no formato dos registros. No entanto, com a expansão e refinamento do corpus, observou-se uma melhora significativa na coerência dos termos sugeridos, tornando a indexação mais alinhada ao conteúdo dos textos analisados.

Além disso, a evolução das pontuações atribuídas aos termos evidenciou um ganho na confiabilidade do modelo, garantindo maior consistência na atribuição dos descritores. Essa melhoria não se limitou a casos isolados, mas foi consistente em todas as publicações testadas, reforçando a importância de um treinamento contínuo e bem estruturado.

A adaptação do *Annif* para publicações em português mostrou-se uma estratégia promissora para automatizar a indexação, reduzindo o esforço manual e garantindo maior padronização na recuperação da informação. O aprimoramento do arquivo de treinamento deve seguir um processo contínuo, incorporando novas fontes de publicações e ampliando o conjunto de dados, o que permitirá um refinamento progressivo dos resultados. Além disso, a experimentação com diferentes *backends* pode aumentar a eficiência da recuperação da informação e contribuir para a normalização dos metadados.

Apesar dos avanços, alguns desafios persistem, como a necessidade de expandir o vocabulário controlado para ampliar a cobertura terminológica e minimizar ambiguidades semânticas. Futuras pesquisas poderão explorar a integração do *Annif* com modelos baseados em redes neurais, aprimorando sua capacidade de generalização e adaptação ao contexto linguístico do português.

REFERÊNCIAS

FUJITA, M. L. Sistema de indexação PRECIS: perspectiva histórica e técnica de seu desenvolvimento e aplicação. **Revista Brasileira de Biblioteconomia e Documentação**, São Paulo, v. 21, n. 1/2, p. 46-62, jan./jun. 1988.

HAYNES, D. **Metadata for information management and retrieval**: understanding metadata and its use. London: Facet Publishing, 2018.

POMERANTZ, J. **Metadata**. Cambridge: MIT Press, 2015.

RUBI, M. P.; FUJITA, M. S. L. Elementos de política de indexação em manuais de indexação de Sistemas de Informação especializados. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 8, n. 1, p. 66-77, jan./jun. 2003. Disponível em: <https://periodicos.ufmg.br/index.php/pci/article/view/23460/18920>. Acesso em: 19 maio 2025.

SUOMINEN, Osmo. Annif: DIY automated subject indexing using multiple algorithms. **Liber Quarterly: the Journal of the Association of European Research Libraries**, [S. l.], v. 29, n. 1, p. 1-26, 2019. Disponível em: <https://doi.org/10.18352/lq.10285>. Acesso em: 19 maio 2025.

SVENONIUS, E. **The intellectual foundation of information organization**. Cambridge, MA: MIT Press, 2000.



[4.0 Internacional](https://creativecommons.org/licenses/by/4.0/)

Esta licença permite compartilhamento, remixe, adaptação e criação a partir do trabalho, mesmo para fins comerciais, desde que sejam atribuídos créditos ao(s) autor(es). Conteúdos elaborados por terceiros, citados e referenciados nesta obra não são cobertos pela licença.

Como citar este trabalho:

COSTA, Bruno Carlos da Cunha; ARAKAKI, Ana Carolina Simionato; NASCIMENTO, Renan Luiz da Silva. Aprimoramento do Annif para indexação automática de publicações em português. In: WORKSHOP DE INFORMAÇÃO DADOS E TECNOLOGIA, 8., 2025, Marília, SP. **Anais [...]**. Marília, SP: Universidade de Marília, 2025. DOI: <https://doi.org/10.22477/viii.widat.273>.



WIDaT 2025

VIII WORKSHOP DE INFORMAÇÃO,
DADOS E TECNOLOGIA

MARÍLIA • SP

Tendências da Agricultura Inteligente: análise das patentes indexadas na base Derwent Innovations Index

Trends in Smart Agriculture: analysis of patents indexed in the Derwent Innovations Index

Janaina Lais Pacheco Lara Morandin¹



<http://lattes.cnpq.br/8844607097133585>



<https://orcid.org/0000-0001-7172-3087>

Fernanda Bochi²



<http://lattes.cnpq.br/8572535542111719>



<https://orcid.org/0000-0002-3275-0725>

Ana Maria Mielniczuk de Moura³



<http://lattes.cnpq.br/1734997653639992>



<https://orcid.org/0000-0002-7811-4660>

Maria Cláudia Cabrini Grácio⁴



<http://lattes.cnpq.br/5170688300970006>



<https://orcid.org/0000-0002-8003-0386>

Resumo

Este estudo utiliza o método patentométrico para analisar patentes em Agricultura Inteligente indexadas na base *Derwent Innovations Index*. A busca recuperou 8.612 patentes usando termos relacionados à Inteligência Artificial e Agricultura Inteligente. Os dados foram organizados em um *dataframe* no *Google Colaboratory* e analisados por meio de *scripts* em *Python*. Para a evolução temática, as patentes foram divididas em três períodos (2016-2018, 2019-2021 e 2022-2024), com redes geradas no UCINET 6. Os resultados indicam crescimento expressivo do patenteamento desde 2018, com a China liderando os depósitos (81,7%). Três áreas concentram 90% das patentes: Necessidades Humanas, Física e Processamento/Transporte. A análise de redes revelou o avanço de tecnologias de Inteligência Artificial, sensores e Internet das Coisas na Agricultura. Estudos futuros podem aprofundar a relação entre inovação

¹ Mestre, Universidade Estadual Paulista Júlio de Mesquita Filho (Unesp), Marília, São Paulo, Brasil. janaina.morandin@unesp.br.

² Doutora, Universidade Estadual Paulista Júlio de Mesquita Filho (Unesp), Marília, São Paulo, Brasil. Pós doutoranda financiada pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) Bolsa nº 88887.081540/2024-00. fernanda.bochi@unesp.br.

³ Doutora, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Rio Grande do Sul, Brasil. Professora Associada no curso de Biblioteconomia, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Rio Grande do Sul, Brasil. ana.moura@ufrgs.br.

⁴ Doutora, Universidade Estadual Paulista Júlio de Mesquita Filho (Unesp), Marília, São Paulo, Brasil. Docente permanente do Programa de Pós-graduação em Ciência da Informação, Universidade Estadual Paulista Júlio de Mesquita Filho (Unesp), Marília, São Paulo, Brasil. cabrini.gracio@unesp.br.

e políticas públicas, assim, sugere-se a análise da correlação entre as patentes de Agricultura Inteligente e os depósitos de patentes na Agricultura convencional.

Palavras-chave: agricultura inteligente; patentometria; Inteligência Artificial; smart agriculture.

Abstract

This study employs the patentometric method to analyze patents in Smart Agriculture indexed in the Derwent Innovations Index database. The search retrieved 8,612 patents using terms related to Artificial Intelligence and Smart Agriculture. The data was organized into a dataframe in Google Colaboratory and analyzed using Python scripts. For thematic evolution, the patents were divided into three periods (2016-2018, 2019-2021, and 2022-2024), with networks generated in UCINET 6. The results indicate significant growth in patent filings since 2018, with China leading the filings (81.7%). Three areas concentrate 90% of the patents: Human Needs, Physics, and Processing/Transport. Network analysis revealed the advancement of Artificial Intelligence technologies, sensors, and the Internet of Things in Agriculture. Future studies could further explore the relationship between innovation and public policies; thus, it is suggested to analyze the correlation between Smart Agriculture patents and patent filings in conventional Agriculture.1414

Keywords: Smart Agriculture; patentometry; Artificial Intelligence; smart agriculture.

1 INTRODUÇÃO

A Agricultura é a maior indústria do mundo, ocupando 50% das terras habitáveis do planeta (Weid, 2022). Assim, para suprir a demanda por alimentos e o desenvolvimento de uma Agricultura estável e sustentável, Cui e Wang (2023) afirmam que a Ciência e a Tecnologia (C&T) são fundamentais, pois permitem acelerar a construção da Agricultura moderna.

Neste contexto, a Agricultura Inteligente aplica Tecnologias da Informação e Comunicações (TIC) ricas em dados, combinadas a *softwares* avançados (Wolfert; Goelse; Sorensen, 2014). As TIC podem incluir o uso de Inteligência Artificial (IA) na implementação, manutenção e uso de processos agrícolas. Assim, às ferramentas convencionais, como tratores e pluviômetros, é adicionado um tipo de IA capaz de executar ações autônomas, ou remotas, fazendo com que o trabalho humano seja cada vez mais auxiliado por máquinas, de forma que o ciclo ciberfísico se torne quase autônomo (Wolfert *et al.*, 2017). Além da incorporação de IA e o uso de *softwares*, a Agricultura Inteligente prevê o uso de *big data*, sensores e drones (Weid, 2022).

No âmbito dos Estudos Métricos da Informação (EMI), por meio da aplicação de indicadores, a Patentometria fornece um panorama do cenário tecnológico,

possibilitando ampliar a compreensão do desenvolvimento tecnológico e contribuindo para a tomada de decisões.

No contexto das vantagens obtidas pelo depósito de patentes e das informações disponíveis nestes documentos, esta pesquisa objetiva analisar as características das patentes em Agricultura Inteligente, indexadas na base *Derwent Innovations Index* (DII). De forma mais específica, objetiva-se caracterizar as patentes quanto a: distribuição por ano de depósito, países depositantes e áreas de classificação, a fim de descrever a evolução do patenteamento relativo à Agricultura Inteligente.

1 REFERENCIAL TEÓRICO

Para Wolfert *et al.* (2017), o conceito de Agricultura Inteligente engloba a aplicação de dispositivos inteligentes conectados à internet, que realizam funções como monitoramento e controle, enquanto auxiliam seres humanos nas tarefas de planejamento e análise. Estes dispositivos permitem o uso de um grande conjunto de dados, sejam os gerados na própria atividade agrícola, quanto os dados externos, como os meteorológicos.

Como fatores que impulsionam o desenvolvimento da Agricultura Inteligente estão, entre outros, as tecnologias baseadas em dados e Internet das Coisas (IoT), a ascensão das empresas de tecnologia agrícola, o sensoriamento remoto, o desenvolvimento de robôs e drones, a geração e o armazenamento de dados e os sistemas abertos de gestão agrícola com aplicativos específicos. Este desenvolvimento promove a melhoria do controle de gestão e tomada de decisões, a segurança alimentar e nutricional, a sustentabilidade, além de prover a quantidade e qualidade de informações (Wolfert *et al.*, 2017).

Nesse cenário, as máquinas são equipadas com todo tipo de sensores que medem uma grande quantidade de dados que são usados para programar o comportamento das máquinas. Estas programações podem ser relativamente simples, ou incluir até mesmo algoritmos de aprendizado profundo (Wolfert *et al.*, 2017).

Dentre as tecnologias desenvolvidas, destaca-se o uso de drones para disseminação de defensivos. Esta prática diminui o uso de aviões agrícolas e a

consequente emissão de poluentes, enquanto uma quantidade menor de produtos químicos é utilizada, em uma área precisamente delimitada. O mercado de drones para a Agricultura na América do Sul foi avaliado em mais de US\$ 9 milhões em 2019 e a projeção é de que alcance mais de US\$ 64 milhões em 2027. Os sensores também são um exemplo importante de tecnologia desenvolvida (Weid, 2022).

De acordo com as inovações inerentes à Agricultura Inteligente, as patentes representam um fator importante no seu desenvolvimento, permitindo proteger as invenções, de forma que o seu titular possa se beneficiar durante um período específico de tempo. Cada patente recebe uma determinada classificação, que permite identificar a tecnologia desenvolvida, dentro de uma área específica. Assim, a evolução de tecnologias pode ser observada por meio das classificações atribuídas ao longo do tempo.

2 PROCEDIMENTOS METODOLÓGICOS

Esta pesquisa utiliza o método patentométrico e a análise de redes de coocorrência para analisar as patentes em Agricultura Inteligente, indexadas na base *Derwent Innovations Index* (DII). A busca pelas patentes foi realizada na base no dia 25/02/2025, aplicando no campo *Topic* (TS) a seguinte estratégia de busca: “(artificial intelligence AND smart farm) OR smart agriculture OR intelligent agriculture”. De acordo com a literatura, os termos se referem à Agricultura moderna, caracterizada pelo uso de IA, mas também, por *softwares* e demais equipamentos tecnológicos. O resultado recuperou 8.612 patentes, exportadas em 9 arquivos de texto sem formatação, com até 1.000 patentes em cada arquivo.

Em seguida, foi criado um *dataframe* usando o *Google Colaboratory*, organizando os seguintes campos recuperados: número da patente, país, título, data, área de classificação e resumo. Os dados foram tratados e exibidos a partir de um *script* em linguagem *Python* que percorre uma lista dos arquivos texto obtidos da base. O *dataframe* foi preenchido executando seleção, limpeza ou contagem, de forma a simplificar as análises. As análises e gráficos foram gerados a partir do *dataframe*, conforme as demandas de observação, utilizando técnicas de sumarização, contagem, agrupamento, similaridade, associação e seleção. Como a base fornece vários números

de classificação de área, optou-se por utilizar o primeiro apontado, assim como o primeiro número da patente, já que os demais se referem à mesma patente, cuja proteção foi requerida em outros locais.

Para a análise de evolução temática as patentes foram divididas em três períodos: de 2016 a 2018, 2019 a 2021 e de 2022 a 2024. Esta divisão permitiu uma melhor visualização das redes geradas. Os anos iniciais, 2014 e 2015 foram descartados, por conterem apenas cinco patentes. As redes foram geradas utilizando o *software* UCINET versão 6. A temática das patentes foi determinada pela conferência da descrição do número de classificação (CIP) disponibilizada pela INPI⁵, versão 2025.1. Quando necessário, versões anteriores da CIP foram consultadas. No caso da descrição da CIP se mostrar insuficiente, foram consultados os títulos e resumos das patentes.

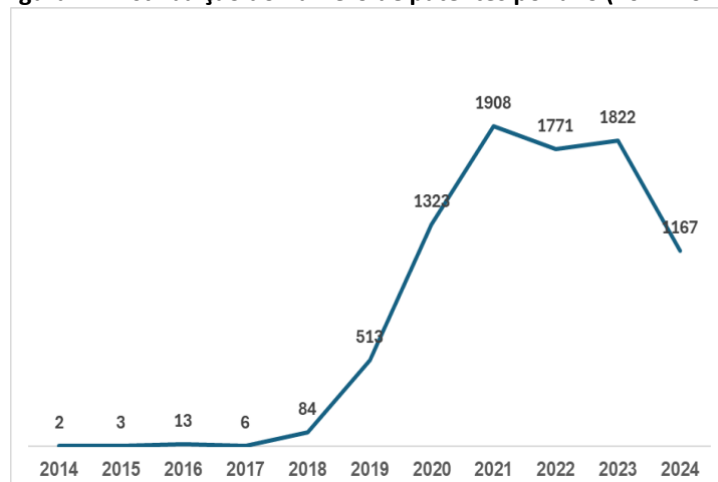
3 ANÁLISE E DISCUSSÃO DOS RESULTADOS

A primeira patente data de 2014, indicando que o patenteamento na temática é recente. A Figura 1 apresenta a distribuição das quantidades de patentes depositadas desde 2014 até 2024, a partir da qual é significativo destacar o ponto de inflexão em 2018, em que houve um aumento de 13 vezes em relação a 2017. Também 2019 configurou um ano de grande aumento da quantidade de patentes, representando 5 vezes o número de patentes depositadas em 2018.

Apesar de se observar o crescimento anual na quantidade de depósitos, com diminuição apenas em 2017 e 2022, os primeiros anos indicam poucas patentes concedidas, com depósito significativo a partir de 2019. A maioria dos depósitos foi realizada em 2021 (22,1%). Destaca-se o fato de uma acentuada queda no número de registros nos anos que correspondem ao período de sigilo de 18 meses ser recorrente para a maioria dos escritórios (2023 e 2024).

⁵ WORLD INTELLECTUAL PROPERTY ORGANIZATION (WIPO). Publicação IPC. Esquema. **SEÇÃO H — ELECTRICIDADE**. 2025. Disponível em: <https://ipc.inpi.gov.br/classifications/ipc/ipcpub/?notion=scheme&version=20250101&symbol=H04L&menulang=pt&lang=pt&viewmode=f&fipcp=no&showdeleted=>. Acesso em: 02 mar. 2025.

Figura 1 - Distribuição do número de patentes por ano (2014-2024)



Fonte: Elaboração própria.

A Tabela 1 apresenta os 24 escritórios ordenados em ordem decrescente de número de patentes depositadas, incluindo o Escritório Europeu de Patentes (EP) e o Tratado de Cooperação em Matéria de Patentes (WO). É possível observar que a China é o país com a maior quantidade de patentes, correspondendo a 81,7%. Os demais países relevantes são Índia, com 10,8% e República da Coreia com 3%. Ademais, cinco escritórios concentram 98% das patentes depositadas no período de 2014 a 2024; a saber: CN, IN, KR, WO e US. Esse resultado vai ao encontro daquele observado por Tey *et al.* (2024), especificamente para patentes em Tecnologias Agrícolas de Precisão. Com dados recuperados da base *Espacenet Patent*, esses autores destacam China, Estados Unidos, Escritório Europeu (EP), Rússia e Coreia como os top 5 países com maior quantidade de patentes depositadas no período de 1960 a 2021.

Tabela 1 – Escritório de depósito das patentes recuperadas

Escritório	Quant. de patentes	%
China (CN)	7038	81.7%
Índia (IN)	928	10.8%
Coreia do Sul (KR)	260	3.0%
Organização Mundial da Propriedade Intelectual (WO)	122	1.4%
Estados Unidos (US)	100	1.2%
Alemanha (DE)	27	0.3%
Austrália (AU)	25	0.3%

Taiwan (TW)	23	0.3%
Escritório Europeu (EP)	16	0.2%
Turquia (TR)	15	0.2%
Japão (JP)	14	0.2%
África do Sul (ZA)	10	0.1%
Luxemburgo (LU)	7	0.1%
Rússia (RU)	7	0.1%
Brasil (BR)	4	0.0%
Romênia (RO)	4	0.0%
Países Baixos (NL)	3	0.0%
Canadá (CA)	2	0.0%
Indonésia (ID)	2	0.0%
Espanha (ES)	1	0.0%
Reino Unido (GB)	1	0.0%
México (MX)	1	0.0%
Malásia (MY)	1	0.0%
República Dominicana (DO)	1	0.0%
Total	8.612	100%

Fonte: Elaboração própria.

De forma geral, o patenteamento agrícola dos Estados Unidos é menor do que na China. No que se refere à Agricultura de Precisão, o estudo de Avola *et al.* (2024) indica que a atividade de patenteamento da China é três vezes maior do que a apresentada pelos Estados Unidos.

Considera-se importante destacar que a China é o maior produtor agrícola mundial. Outros oito países presentes na Tabela 1 também estão entre os 10 maiores produtores agrícolas do mundo (Agrofy News, 2025). Esse fato evidencia a importância da agricultura para esses países e a consequente necessidade de proteção das inovações tecnológicas em seus territórios.

O Brasil possui quatro patentes recuperadas na DII. Porém, estudo conduzido pelo INPI em 2022, sobre as patentes em Agricultura Inteligente depositadas no Brasil, indica que o país é o segundo maior depositante, com 783 pedidos. Segundo o INPI, este número pode ser subestimado, já que as patentes nesta área envolvem, muitas vezes, o desenvolvimento de *softwares*, que no Brasil não são protegidos por patentes (Weid, 2022). É significativo observar que o Brasil apresenta padrão internacional de publicações científicas na área de Agricultura, com um sistema de pesquisa composto pela Empresa Brasileira de Pesquisa Agropecuária (Embrapa), universidades e instituições estaduais de pesquisa (Vanz *et al.*, 2023). Esse padrão pode ser constatado por seu grande volume de produção científica, ocupando, segundo o Ranking ScimagoJR⁶, o terceiro país com maior quantidade de artigos publicados no período de 2014 e 2021 e quarto país mais produtivo na área nos anos de 2022 e 2023. Desse modo, é possível inferir que a ciência gerada não tem convergido de forma expressiva para o patenteamento de novas tecnologias, sugerindo uma desintegração no sistema de Ciência, Tecnologia e Inovação (CTeI) do país.

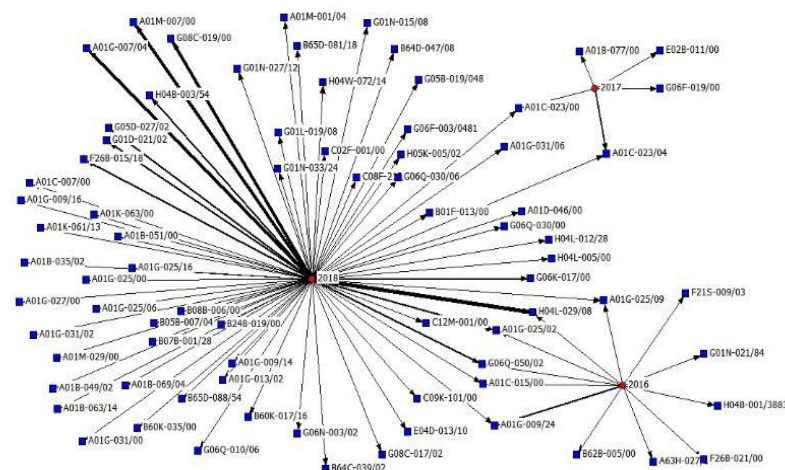
Na distribuição da classificação das patentes se observam as oito áreas da CIP, proposta pela *World Intellectual Property Organization* (WIPO). Três áreas concentram quase 90% dos depósitos: Necessidades Humanas (A) com 46.07%, Física (G) com 33,5% e Operações de Processamento; Transporte (B) com 9,6%. As demais áreas demonstram menor expressividade: Eletricidade (H) com 5,1%, Armas; Explosão (F) com 3.55%, Química; Metalurgia (C) com 1.14%, Construções fixas (E) com 1.02% e Papel; Têxteis (D) com 0.02%.

A área de Necessidades Humanas, na CIP, especificamente na classe 01 trata da Agricultura. Assim, é esperado que a maioria das patentes esteja classificada nesta área. Na área da Física (G), estão, normalmente, as patentes que utilizam alguma forma de IA no seu método ou processo, o que justifica essa ser a segunda área com maior número de registros.

⁶ SCIMAGO JOURNAL and COUNTRY RANK (SJR). **Home**. c2025. Disponível em: <https://www.scimagojr.com/>. Acesso em: 2 mar. 2025.

A rede de coocorrência formada pelo primeiro período, de 2016 a 2018, indica a predominância de patentes classificadas na área de Necessidades Humanas (A), conforme a Figura 2. Este é o período com a menor incidência de patentes.

Figura 2 – Rede de Coocorrência de Classificações de Patentes (2016-2018)



Fonte: Elaboração própria.

As classificações destacadas, cujas arestas são representadas mais grossas na rede, se referem a duas patentes classificadas na área de Necessidades Humanas (A), tanto em 2016 quanto em 2017. Em 2018, 13 classificações se destacam, com ênfase no número H04L-029/08 que se repete quatro vezes.

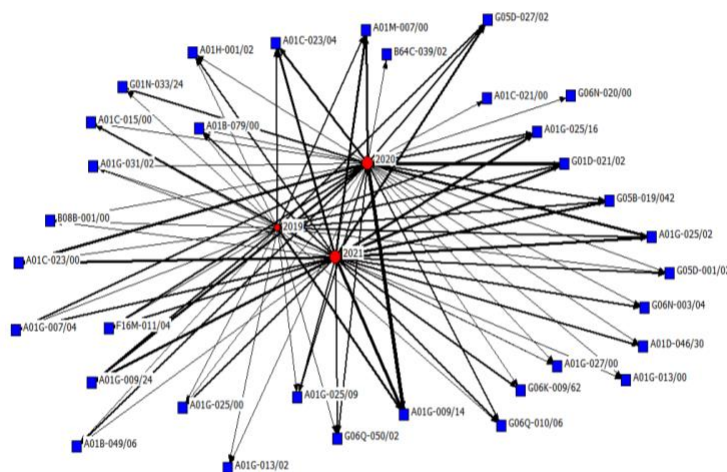
Estas classificações descrevem métodos que utilizam dispositivos, controle de transmissão de dados e sistemas de transmissão de sinais elétricos. A classificação A01G-007/04, que possui três patentes em 2018, trata de uma nova tecnologia referente à bioestimulação elétrica e magnética, que visa aumentar a produtividade agrícola de forma sustentável. Os dados controlados podem ser empregados na tripulação remota, evitando colisões de máquinas e também no controle de fluxo, evitando que sejam transmitidos mais dados do que um receptor possa processar.

A classe A01M-007/00, com três patentes em 2018, é empregada nas invenções que utilizam a IA. É perceptível a relevância das classificações destacadas no triênio, pois refletem os objetivos da Agricultura moderna, como a produtividade maximizada e sustentável.

É possível observar, na rede, classificações compartilhadas entre os anos de 2016 e 2018 e entre 2017 e 2018, indicando que a tecnologia patenteada é continuada no período. Os anos de 2016 e 2018 compartilham quatro classificações, enquanto 2017 e 2018 compartilham duas. A classificação H04L-029/08 destacada em 2018, é comum a 2016, indicando que os processos de controle de dados são tecnologias presentes desde o início do patenteamento.

A Figura 3 ilustra a coocorrência de classificações entre 2019 e 2021. Este período detém 43,4% das patentes.

Figura 3 – Rede de Coocorrência de Classificações de Patentes (2019-2021)



Fonte: Elaboração própria.

No triênio de 2019 a 2021, observou-se uma mudança significativa nas menções à Classificação Internacional de Patentes (CIP). Durante esse período, houve um aumento no número de classificações citadas, sendo a maioria mencionada apenas uma vez em um dos três anos analisados. Para a análise da rede, foram consideradas apenas as classes com mais de uma menção e que apareceram em pelo menos dois anos. Nesse contexto, destacaram-se 26 classes mencionadas nos três anos do período, todas com mais de uma ocorrência. Dentre elas, 14 pertencem à seção A (Necessidades Humanas), uma à seção B (Operações de Processamento; Transporte), duas à seção F (Armas; Explosão), oito à seção G (Física) e uma à seção H (Eletricidade), especificamente a H04L-029/08, que foi mencionada duas vezes em 2019, três vezes em 2020 e 15 vezes em 2021, sendo representativa também no triênio de 2016-2018.

A classificação com maior representatividade é a A01G-009/14, mencionada 37, 88, 69 respectivamente nos três anos. Além dela, destacam-se nas Necessidades Humanas as classificações A01G-009/24 (26, 41, 51), A01G-025/02 (17, 48, 43) e A01M-007/00 (15, 39, 50). Outras classificações que ocorreram com muita expressividade nos anos de 2019 a 2021, foram G01D-021/02 (22, 56, 45), G05B-019/042 (13, 16, 43) e a G05D-027/02 (17, 24, 39).

Nestes anos, as classificações remetem a patentes que desenvolveram sensores utilizados, por exemplo, na otimização da irrigação. Observa-se o uso da IA, com aplicação em drones e na análise de imagens. As classificações indicam patentes que utilizam aparelhos de radiofrequência como a tecnologia *Low-Rank Adaptation* (LoRA), que utiliza pouca energia para transmissão de baixo volume de dados a longas distâncias, em locais de baixa infraestrutura, como fazendas. Observou-se, também, o uso de tecnologias baseadas na IoT.

A rede de coocorrência de 2022 a 2024 (Figura 4), mostra o período com maior incidência de patentes, 55,2%. Na rede, entre os anos de 2022 a 2024, identificaram-se 13 classificações mencionadas consecutivamente ao longo dos três anos, todas com mais de 10 ocorrências.

Dentre essas classificações, nove pertencem à seção A (Necessidades Humanas) e quatro à seção G (Física). São elas, respectivamente, A01B-079/00 (13, 29, 35), A01C-023/00 (52, 48, 17), A01C-023/04 (51, 54, 15), A01G-009/14 (56, 42, 12), A01G-009/24 (45, 39, 12), A01G-013/02 (17, 11, 10), A01G-025/09 (23, 23, 16), A01M-007/00 (38, 30, 14). Dentro das Necessidade Humanas, destacou-se a classificação A01G-025/16, que apresentou menções acima de 50 nos três anos (54 menções em 2022, 81 em 2023 e 73 em 2024). Na Física (G) o destaque ficou para G01D-021/02 (49, 37, 23).

As tecnologias patenteadas fazem menção, principalmente, à transmissão de sinal e dados em tempo real. Observa-se também a presença de aprendizado de máquina, mas, sobretudo, predomina a IoT, com o uso de sensores e câmeras. Pode-se considerar que as classificações deste triênio indicam as tendências no patenteamento em Agricultura Inteligente a nível mundial.

produção agrícola de forma sustentável. Conclui-se que o patenteamento é recente, iniciando em 2014, com crescimento marcante em 2019. Dentre os depositantes, a China é o país que possui a grande maioria dos depósitos, seguida por Índia e República da Coreia. A área com a maior quantidade de patentes é Necessidades Humanas, seguida pela área da Física.

A evolução temática, observada pelas classificações da CIP, mostrou que as tecnologias iniciadas em 2016 continuaram a ser desenvolvidas até 2018, com foco no aprimoramento de maquinário e da automatização de ações ligadas ao cultivo e ao solo. O segundo triênio apontou um aumento no uso da IA, o que se mantém no terceiro período. Desde o início, observou-se a predominância de patentes voltadas à transmissão de dados, com patentes utilizando sensores ligados a redes. Assim, percebe-se o aumento do uso de IA nas invenções, porém, a Internet das Coisas é o conceito tecnológico mais empregado em todos os períodos analisados, representando a maior tendência no patenteamento em Agricultura Inteligente.

Sugere-se, como estudos futuros, relacionar as patentes em Agricultura Inteligente ao depósito de patentes na Agricultura convencional. Propõem-se também, a relevância da investigação da interação que envolve as produções científicas e as patentes depositadas pelos países, investigando-se a relação entre Ciência, Tecnologia e Inovação.

REFERÊNCIAS

AGROFY NEWS. Quem são os maiores produtores agrícolas do mundo? **Agrofy News**, 29 jun. 2023. Disponível em: <https://news.agrofy.com.br/noticia/201932/quem-sao-os-maiores-produtores-agricolas-do-mundo>. Acesso em: 05 mar. 2025.

AVOLA, G.; DISTEFANO, M.; TORRISI, A.; RIGGI, E. Precision agriculture and patented innovation: state of the art and current trends. **World Patent Information**, [S. l.], v. 76, e102262, 2024. DOI: <https://10.1016/j.wpi.2024.102262>. Disponível em: <https://doi.org/10.1016/j.wpi.2024.102262>. Acesso em: 21 fev. 2025.

CUI, L.; WANG, W. Factors affecting the adoption of digital technology by farmers in China: a systematic literature review. **Sustainability**, [S. l.], v. 15, n. 20, e14824, 2023. DOI: <https://10.3390/su152014824>. Disponível em: <https://doi.org/10.3390/su152014824>. Acesso em: 12 maio 2025.

TEY, Y. S.; BRINDAL, M.; WONG, S. Y.; IBRAGIMOV, A. A.; YUSOP, M. R. Evolution of precision agricultural technologies: a patent network analysis. **Precision Agriculture**, [S. l.], v. 25, p. 376-395, 2024. DOI: <https://10.1007/s11119-023-10076-y>. Disponível em: <https://doi.org/10.1007/s11119-023-10076-y>. Acesso em: 05 mar. 2025.

VANZ, S. A. S.; GRACIO, M. C. C.; OLIVEIRA, S. C.; CHINCHILLA-RODRÍGUEZ, Z.; DOCAMPO, D. Collaboration strategies and corresponding authorship in agronomy research of brazilian academic and non-academic institutions. **Scientometrics**, [S. l.], v. 128, p. 6403-6426, 2023. DOI 10.1007/s11192-023-04857-5. Disponível em: <https://doi.org/10.1007/s11192-023-04857-5>. Acesso em: 27 nov. 2024.

WEID, I. V. D. **Radar tecnológico**: análise do patenteamento de tecnologias relacionadas à agricultura sustentável depositadas no Brasil. Rio de Janeiro: INPI, 2022. 53 p. Disponível em: https://www.gov.br/inpi/pt-br/assuntos/informacao/RadarAgriculturaSustentavel_final_2022.pdf. Acesso em: 23 fev. 2025.

WOLFERT, S.; GE, L.; VERDOUW, C.; BOGAARDT, M. J. Big data in smart farming: a review. **Agricultural Systems**, [S. l.], v. 153, p. 69-80, may 2017. Disponível em: <https://doi.org/10.1016/j.agsy.2017.01.023>. Acesso em: 02 mar. 2025.

WOLFERT, S.; GOENSE, D.; SORENSEN, C. A. G. A future internet collaboration platform for safe and healthy food from farm to fork. In: ANNUAL SRII GLOBAL CONFERENCE, 2014, San Jose, CA. **Conference [...]**. San Jose: IEEE, 2014. p. 266-273. Disponível em: <https://ieeexplore.ieee.org/document/6879694>. Acesso em: 27 nov. 2024.



[4.0 Internacional](https://creativecommons.org/licenses/by/4.0/)

Esta licença permite compartilhamento, remixe, adaptação e criação a partir do trabalho, mesmo para fins comerciais, desde que sejam atribuídos créditos ao(s) autor(es). Conteúdos elaborados por terceiros, citados e referenciados nesta obra não são cobertos pela licença.

Como citar este trabalho:

MORANDIN, Janaina Lais Pacheco Lara; BOCHI, Fernanda; MOURA, Ana Maria Mielniczuk de; GRÁCIO, Maria Cláudia Cabrini. *Tendências da AGRICULTURA INTELIGENTE: análise das patentes indexadas na base Derwent Innovations Index*. In: WORKSHOP DE INFORMAÇÃO DADOS E TECNOLOGIA, 8., 2025, Marília, SP. **Anais [...]**. Marília, SP: Universidade de Marília, 2025. DOI: <https://doi.org/10.22477/viii.widat.258>.



WIDaT 2025

VIII WORKSHOP DE INFORMAÇÃO,
DADOS E TECNOLOGIA

MARÍLIA • SP

Algoritmo para análise de presença *online* de pesquisadores da base ORCID

Algorithm for analyzing the online presence of researchers in the ORCID database

Rafael Gutierrez Castanha¹



<http://lattes.cnpq.br/4834832439175113>



<https://ORCID.org/0000-0002-3117-1780>

Francielle Franco²



<http://lattes.cnpq.br/0132426572975221>



<https://ORCID.org/0000-0002-3686-595X>

Resumo

Este estudo teve como objetivo desenvolver um algoritmo capaz de identificar automaticamente a presença *online* de pesquisadores registrados na ORCID, analisando websites, mídias sociais e outros identificadores acadêmicos vinculados aos perfis. Para isso, foi desenvolvido um código em R que acessa a API pública da ORCID, extrai informações dos perfis cadastrados e gera um relatório com quatro tabelas (dados gerais, dados sumarizados, País x Plataforma, Coocorrência entre Plataformas) e um grafo de coocorrência entre as plataformas cadastradas. A prova de conceito foi realizada com 997 pesquisadores de universidades brasileiras, resultando na geração de quatro tabelas e uma rede coocorrência entre as plataformas. Os resultados demonstraram que 37,2% dos pesquisadores analisados possuíam informações sobre presença *online*, ou seja, este percentual de pesquisadores inseriu os links de seus perfis em outras plataformas nos campos determinados dentro da ORCID. As plataformas mais frequentes foram o *Currículo Lattes*, *Linkedin* e *Google Scholar*. A análise evidencia o potencial da ORCID como agregador de presença *online*, mas ressalta a importância da atualização dos perfis pelos pesquisadores. Como desenvolvimento futuro, propõe-se a criação de uma interface gráfica para facilitar o uso do algoritmo.

Palavras-chave: dados científicos; mídias sociais; identificadores únicos; ORCID; presença *online*.

Abstract

This study aimed to develop an algorithm capable of automatically identifying the online presence of researchers registered on ORCID, by analyzing websites, social media platforms, and other academic identifiers linked to their profiles. To achieve this, an R script was developed to access ORCID's public API, extract information from registered profiles, and generate a report with four tables (general data, summarized data, Country x Platform, and Platform Co-occurrence) as well as a co-occurrence network graph of the registered platforms. The proof of concept was carried out with 997 researchers from Brazilian universities, resulting in the generation of the four tables and a co-occurrence network among platforms. The results showed that 37.2% of the analyzed researchers had information regarding their online presence, meaning this percentage had inserted links to their profiles on other platforms in the designated fields within ORCID. The most frequently mentioned platforms were the Lattes Curriculum, LinkedIn, and Google Scholar. The analysis highlights the potential of ORCID as an aggregator of

¹ Doutor em Ciência da Informação, Universidade Estadual Paulista (Unesp), Marília, São Paulo, Brasil. Docente, Universidade de Marília (Unimar), Marília, São Paulo, Brasil. r.castanha@gmail.com.

² Mestre em Ciência da Informação, Universidade Federal do Rio Grande do Sul (UFRGS), Rio Grande do Sul, Brasil. Doutoranda em Ciência da Informação, Universidade Estadual Paulista (Unesp), Marília, São Paulo, Brasil. francielle.franco@unesp.br.

online presence, while also emphasizing the importance of researchers keeping their profiles up to date. As a future development, the creation of a graphical user interface is proposed to facilitate the use of the algorithm.

Keywords: scientific data; social media; unique identifiers; ORCID; *online* presence.

1 INTRODUÇÃO

A presença *online* de pesquisadores refere-se à forma como os pesquisadores são percebidos no espaço digital, podendo influenciar significativamente sua reputação profissional e impacto acadêmico (Araújo; Pedri, 2023). Essa presença é crucial pois permite que os pesquisadores compartilhem seus trabalhos, se conectem com outras pessoas e participem de conversas científicas mais amplas: A participação ativa em discussões *online* e o compartilhamento de pesquisas podem levar a um maior reconhecimento e oportunidades de colaboração (Bik; Goldstein, 2013). Uma das formas de medir a eficácia dessa presença é por meio da autoridade social, que reflete sua influência e alcance em plataformas como o X por exemplo (Araújo; Pedri, 2023). Portanto, uma presença on-line robusta é essencial para que os pesquisadores estabeleçam credibilidade e promovam a colaboração em suas áreas (Bik; Goldstein, 2013; Araújo; Pedri, 2023).

A importância de pesquisadores terem perfis em plataformas da web é cada vez mais reconhecida na comunidade acadêmica, pois podem aumentar a visibilidade, o engajamento e a reputação profissional. Ter perfis ativos e atualizados em plataformas como o *ResearchGate* ou *Google Scholar* é importante para que um pesquisador possa mostrar o seu trabalho, o que pode levar a uma maior visibilidade dentro da comunidade acadêmica (Orduña-Malea; Martín-Martín; Delgado-López-Cózar, 2016). Entretanto, pensando em termos de divulgação científica, destaca-se que a integração com as mídias sociais é fundamental porque contribui para a reputação *online* de um pesquisador e facilita o engajamento com um público mais amplo (Araújo, 2014). Portanto, manter perfis em plataformas da web é crucial para que pesquisadores aumentem sua visibilidade, promovam conexões profissionais e meçam seu impacto no cenário acadêmico.

Dentre as plataformas que disponibilizam a criação de perfis de pesquisadores destacam-se aquelas de cunho acadêmico (*ResearchGate* e *Google Scholar*), as de uso pessoal e profissional (*Facebook*, *X*, *Linkedin*) e os perfis que disponibilizam identificadores únicos a parte do registro em base de dados (ORCID, Researcher iD e Scopus Author iD). No que tange

ao aprimoramento e estabelecimento da presença *online* de pesquisadores em distintos ambientes *online* observa-se que a plataforma ORCID (*Open Researcher and Contributor ID*) tem papel fundamental (Orduña-Malea; Martín-Martín; Delgado-López-Cózar, 2016).

Nesse sentido, a gestão de identidade acadêmica refere-se ao conjunto de práticas e estratégias adotadas por pesquisadores para organizar, manter e divulgar suas informações profissionais, produção científica e conexões institucionais em ambientes digitais. Essa gestão envolve a atualização constante de perfis em plataformas acadêmicas (como ORCID, *ResearchGate*, *Google Scholar*, entre outras), bem como o uso consciente das mídias sociais como ferramentas de visibilidade e reputação profissional (Barata *et al.*, 2018). Costas, Corona e Robinson-García (2024) destacam que a manutenção de perfis atualizados em plataformas como o ORCID tem papel central no ecossistema da pesquisa, por integrar informações dispersas e promover maior transparência e interoperabilidade na comunicação científica. Assim, a gestão da identidade acadêmica vai além da curadoria pessoal de dados, sendo uma prática essencial para o fortalecimento da presença digital dos pesquisadores e para a integridade dos sistemas de avaliação e análise da ciência.

Para além da função de garantir a correta atribuição de autoria, a ORCID ainda pode ajudar os pesquisadores a manterem uma identidade *online* consistente, vinculando seu trabalho em várias plataformas (Costas; Corona; Robinson-García, 2024). O fornecimento de um identificador único facilita o gerenciamento centralizado de perfis, integrando-se a outras plataformas, aumentando a visibilidade e incentivando atualizações regulares (Wang *et al.*, 2024). Assim, ter um perfil na ORCID não apenas ajuda os pesquisadores individuais a gerenciar suas atividades acadêmicas, mas também melhora a qualidade geral e a confiabilidade dos dados acadêmicos.

A ORCID foi projetada para atender as necessidades de pesquisadores, agências de fomento e instituições e esse aspecto comunitário solidifica ainda mais seu papel como agregador, pois conecta várias partes interessadas e facilita o compartilhamento de informações (Costas; Corona; Robinson-García, 2024). Embora a ORCID tenha o potencial de atuar como um agregador de presença *online* para pesquisadores, sua eficácia é influenciada pela integridade dos perfis e pelas motivações por trás de sua adoção. Assim, o aprimoramento da atualização dos perfis ORCID poderia solidificar ainda mais sua posição

como uma ferramenta fundamental para a manutenção da presença *online* de pesquisadores (Arroyo-Machado *et al.*, 2025; Wang *et al.*, 2024).

Os perfis na ORCID viabilizam um repositório estruturado de informações de pesquisadores, facilitando a integração de dados e ressaltando sua importância como fonte de dados para várias atividades relacionadas à pesquisa (Costas; Corona; Robinson-García, 2024). A ORCID conta com mais de 9 milhões de cadastros, sendo o Brasil o terceiro país com maior número de perfis (447.181 mil). Quanto à integridade dos dados, destaca-se que 66% dos perfis apresentam a informação de filiação, 50% apresentam informações de publicações. Quando esses dados são analisados em conjunto, identifica-se que apenas 34% dos perfis têm ambas as informações registradas. Existem diversas possibilidades de integração com a plataforma, atualmente a ORCID conta com 5941 integrações ativas (ORCID, 2025).

Arroyo-Machado *et al.* (2025) coletaram os metadados dos perfis de pesquisadores espanhóis para estudar a frequência das atualizações, bem como investigar seções de metadados preenchidas e características comuns entre os perfis e por fim aumentar a confiabilidade da ORCID para rastrear a mobilidade do pesquisador. Costas, Corona, Robinson-García (2024) também relacionaram os dados da ORCID como um facilitador na análise das relações de financiamento e da mobilidade de pesquisadores. Enquanto base de dados, a ORCID é explorada em função de seu potencial para analisar aspectos relacionados à carreira acadêmica no Brasil, incluindo a mobilidade entre instituições (Carneiro; Pinto; Gomes, 2024). Barata *et al.* (2018) discute que a integração da ORCID às práticas acadêmicas no Brasil faz parte de um movimento maior para melhorar a presença *online* dos pesquisadores, impulsionado pelas políticas institucionais e pelo cenário em evolução da comunicação científica.

Portanto, a crescente importância da presença *online* dos pesquisadores na web, evidenciada pelo impacto das mídias sociais, destaca a necessidade de mecanismos automatizados para identificar e analisar essa presença. Assim, define-se como questão norteadora: de que maneira um algoritmo pode contribuir para a extração e visualização de informações sobre a presença *online* de pesquisadores registrados na ORCID, considerando websites, mídias sociais e outros identificadores cadastrados? O objetivo deste estudo é desenvolver um algoritmo capaz de identificar automaticamente a presença *online* de

pesquisadores registrados na ORCID, analisando websites, mídias sociais e outros identificadores acadêmicos vinculados aos perfis.

2 PROCEDIMENTOS METODOLÓGICOS

A fim de construir um algoritmo capaz de identificar a presença *online* de pesquisadores registrados na ORCID, desenvolveu-se um algoritmo, em linguagem de programação R, capaz de acessar os dados cadastrais informados pelos pesquisadores em seus respectivos perfis na ORCID. O código é aberto e foi disponibilizado via repositório digital Zenodo em Castanha e Franco (2025). Para acessar estes dados, utilizou-se a API (*Application Programming Interface*) pública da base de dados da ORCID, que permite a consulta aos dados de um perfil de pesquisador que tenha tornado as suas informações públicas. As respostas das consultas via API são em formato *Extensible Markup Language* (XML).

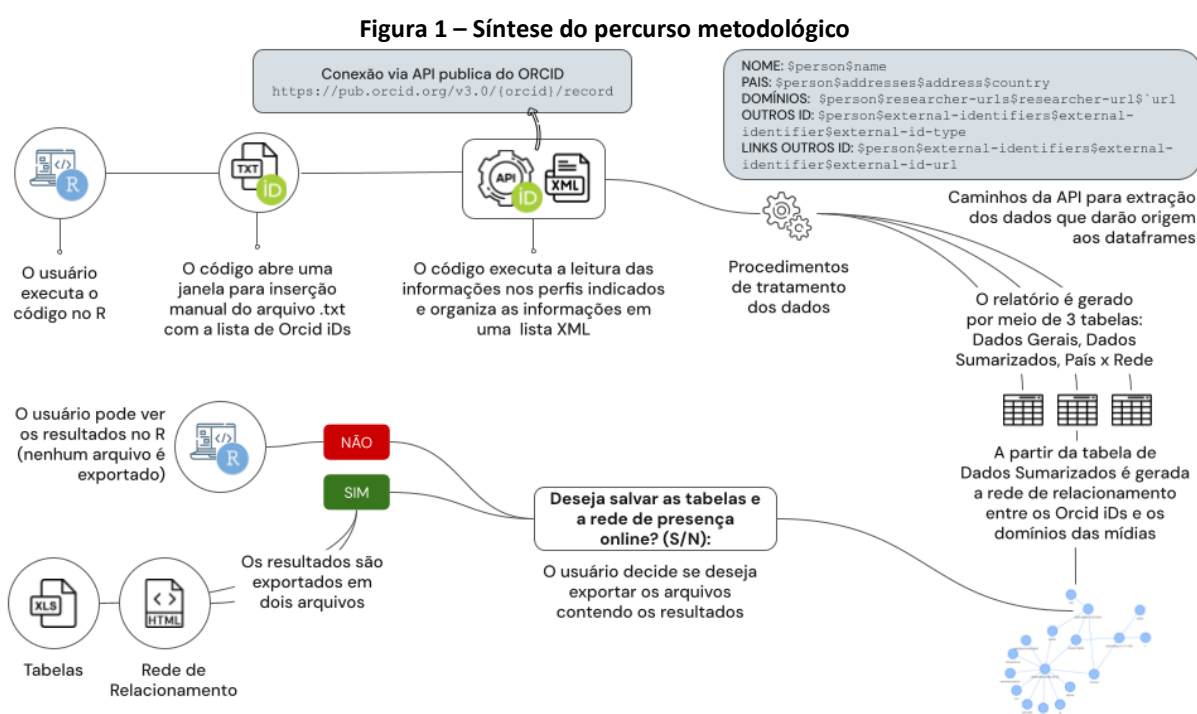
Tendo foco na presença *online* dos pesquisadores registrados na base, o algoritmo em questão deveria ser capaz de coletar os dados referentes a: nome de cada pesquisador, país de origem, nomes e *urls* dos websites e mídias sociais (*Websites e social links*), nomes e *urls* de outras plataformas fornecedoras de identificadores únicos (*Other IDs*). Para que essas informações fossem coletadas, definiu-se a necessidade de submeter uma lista de números de ORCID IDs ao algoritmo para que o código fosse capaz de identificar e extrair todas as informações requeridas. Assim, o usuário necessita enviar ao algoritmo uma lista contendo uma única coluna com os ORCID IDs de interesse em seu formato alfanumérico padrão, contendo 16 caracteres (0000-0000-0000-000X), sem a parte da *url* que precede a parte numérica (<https://ORCID.org/>). Portanto, é pré-requisito para uso deste código que o usuário possua uma lista de ORCID IDs inseridos em um arquivo .txt, organizados conforme descrito.

Tendo a lista de ORCID IDs, identificou-se na API os campos e os caminhos das informações (supracitadas) a serem extraídas. Estas informações encontram-se nos seguintes caminhos: ORCID IDs:

- a) ORCID IDs: `$`ORCID-identifier`$path`
- b) Primeiro Nome: `$person$name$`given-names``;
- c) Sobrenome: `$person$name$`family-name``
- d) País de origem: `$person$addresses$address$country`

- e) Websites & social links: `$person$`researcher-urls`$`researcher-url`$`url-name``
- f) Links dos websites e mídias sociais: `$person$`researcher-urls`$`researcher-url`$`url``
- g) Outros Identificadores: `$person$`external-identifiers`$`external-identifier`$`external-id-type``
- h) Links dos outros Identificadores: `$person$`external-identifiers`$`external-identifier`$`external-id-url``

Na etapa seguinte realizou-se a extração de todas as informações. Para fins de organização foi gerado um relatório contendo todos os dados obtidos a partir de cada perfil da ORCID. Este relatório é composto por quatro tabelas e uma rede (grafo) que relaciona as plataformas cadastradas. Todo processo é ilustrado na Figura 1.



Fonte: Elaboração própria.

A primeira tabela refere-se aos dados gerais de cada pesquisador contendo oito colunas: Número de ORCID (ORCID IDs), Nome, país de origem, websites e mídias sociais (*Websites and social links*), *urls* dos respectivos websites e mídias sociais, Domínio (Plataforma), outros identificadores (*Other IDs*) e as respectivas *urls* dos outros identificadores. O item descrito como “Domínio (Plataforma)” foi gerado a partir das *urls* de

cada website e mídia social referindo-se aos seus respectivos domínios, por exemplo, em casos que o pesquisador registrou *Google Scholar Profile*, *Google Scholar*, *Google Acadêmico*, ou similares, estes rótulos foram armazenados na coluna “Web sites e social links”. A partir das *url* destes rótulos (presente na quarta coluna), extraiu-se somente o domínio de cada *url*. Por exemplo, todos rótulos exemplificados possuem o mesmo domínio que é *https://scholar.google.com/{...}*. Com isso, o código exclui todos os elementos antes e depois de “scholar”, sendo esta a informação registrada na coluna “Domínio (Plataforma)”. Este procedimento é realizado a partir da criação de um Regex. Esta operação foi primordial para padronização das mídias sociais de cada pesquisador, visto que, diversos rótulos de mídias sociais podem corresponder a um mesmo domínio.

A segunda tabela gerada pelo algoritmo é uma tabela sumarizada contendo somente cinco colunas: Nome, número de ORCID (ORCID IDs), país de origem, domínio e Outros ID. Esta tabela possui linhas com valores únicos, aglutinando nas colunas “Domínio (Plataforma)” e “Outros ID”, em uma única célula, estas informações utilizando o separador ponto e vírgula (;). Esta tabela resume o panorama da presença *online* do conjunto de pesquisadores a serem analisados.

A terceira aba apresenta uma tabela conjugada de frequência, refere-se ao cruzamento de dados entre cada domínio do website ou mídia social e o país de origem dos pesquisadores. Por fim a tabela 4 apresenta os dados de coocorrência entre Plataformas, ou seja, com qual frequência as plataformas (expressas pelos domínios) são mencionadas simultaneamente em cada cadastro. A partir da geração destas quatro tabelas, o algoritmo fornece a opção de salvá-las em arquivo externo de extensão *.xlsx (Excel)* gerando quatro abas distintas: Aba 1: Dados Gerais; Aba 2: Dados Sumarizados; Aba 3: País x Plataforma; Aba 4: Coocorrência entre Plataformas.

O algoritmo ainda gerou uma rede (grafo) de relacionamento entre as plataformas mais mencionadas de forma simultânea em cada cadastro. Para isso, utilizou-se a biblioteca *visnetwork3D*. A rede gerada fornece uma visão geral da presença *online* de pesquisadores em websites e mídias sociais, de modo que cada aresta da rede denota um relacionamento de presença *online* de um pesquisador específico em diferentes plataformas. A opção pelo uso da rede de coocorrência se dá pela possibilidade de visualizar como websites e mídias sociais aparecem conjuntamente nos perfis de pesquisadores registrados na ORCID. A partir de um

conjunto de ORCID IDs analisados, essa abordagem permite identificar padrões de conexão, similaridades na presença *online* e a distribuição das atividades dos pesquisadores em diferentes plataformas.

Cabe ressaltar que tanto o processo de inserção, quanto a exportação dos dados ocorre a partir da interação do usuário com o código. Quando o código é executado, o console no R apresenta a frase “Selecione seu conjunto de ORCID IDs”, abre-se uma janela para a inserção manual de arquivo .txt. Após essa ação, o console mostra a mensagem “Aguarde...”. Passado o processamento, o console exibe a mensagem: “Deseja salvar as tabelas e a rede de presença *online*? (S/N):”, caso o usuário digite “S” ele confirma a exportação do relatório (tabelas) e a rede de presença *online* no formato .html. Desse modo os arquivos são exportados e salvos no diretório padrão do usuário.

A fim de testar o algoritmo proposto, realizou-se uma prova de conceito para demonstração das funcionalidades do programa. Realizou-se a coleta de um conjunto de 1000 ORCID IDs composto pelos 200 primeiros resultados de busca no site da ORCID (ORCID.org) das cinco universidades brasileiras mais bem ranqueadas no *Scimago Institution Ranking* 2025: Universidade de São Paulo (USP), Universidade de Campinas (UNICAMP), Universidade Estadual Paulista (UNESP), Universidade Federal de Minas Gerais (UFMG) e Universidade Federal do Rio de Janeiro (UFRJ). Buscou-se pela sigla de cada uma das universidades para gerar o conjunto com os 200 primeiros resultados de cada busca, gerando uma lista de 1000 ORCID IDs. Este conjunto de dados, bem como os resultados gerados pelo algoritmo também estão no repositório digital Zenodo em Castanha e Franco (2025).

3 RESULTADOS

A partir da prova de conceito, foi possível testar as funcionalidades propostas pelo algoritmo. Sendo assim, foram coletadas todas as informações de 997 pesquisadores únicos (3 registros estavam duplicados) vinculados às referidas universidades. O algoritmo coletou as informações gerais de cada pesquisador, bem como, foi capaz de identificar os perfis de pesquisadores que não cadastraram qualquer informação nos campos consultados. Desse modo, pode-se analisar a presença *online* de 371 pesquisadores, isso porque 626 foram identificados como “Sem presença *online*”, ou seja, não apresentavam qualquer informação

nos campos correspondentes aos websites, mídias sociais ou outros identificadores únicos. Esta ausência pode indicar limitações no preenchimento dos perfis pelos próprios usuários. Tal constatação reforça as discussões sobre a importância da motivação pessoal dos pesquisadores para manterem seus perfis atualizados visando maior eficácia da ORCID como agregador de presença *online* (Arroyo-Machado *et al.*, 2025; Sixto-Costoya *et al.*, 2021).

Ao observar as tabelas geradas pelo algoritmo, verificou-se que o campo “Nome” apresentou algumas inconsistências sendo elas: um pesquisador que não preencheu o nome, oito perfis que fizeram uso dos prefixos “Professor” ou “Doutor”, e ainda oito perfis que não correspondiam à pesquisadores individuais mas sim à autorias institucionais, como revistas, fóruns ou portais de periódicos, como por exemplo: Fórum de Pesquisa Discente FORPED/PPGGOC/UFMG (0000-0003-3627-6482), UFMG Revista Docência do Ensino Superior (0000-0002-2372-1914), Portal de Periódicos UFMG (0000-0003-4190-1430). Sixto-Costoya *et al.* (2021) evidenciaram que a falta de uniformidade nos dados do ORCID dificulta a extração e interpretação automática das informações, além de representar um obstáculo para estudos que dependem de dados sobre a atividade científica e a identidade dos pesquisadores. Portanto, a falta de padronização no preenchimento dos dados de identificação dos pesquisadores pode ter impacto nas análises subsequentes.

Referente à distribuição geográfica dos pesquisadores observou-se que, para além do Brasil (879 pesquisadores únicos), outros países foram identificados em menores proporções como Ucrânia (15 pesquisadores), Estados Unidos (12 pesquisadores) e Espanha (10 pesquisadores). O cruzamento das informações de País X Plataforma (na tabela de dados sumarizados) permitiu a análise da frequência das plataformas em cada país. Destaca-se que no Brasil foram identificadas 93 plataformas diferentes (perfis em mídias sociais, sites institucionais, plataformas fornecedoras de IDs, páginas pessoais, etc.), sendo o *Currículo Lattes* (151 ocorrências), o *Linkedin* (32 ocorrências), o *Instagram* (12 ocorrências) e o *ResearchGate* (9 ocorrências) as plataformas mais mencionadas. Portanto, o algoritmo proposto pode apresentar dados de presença *online* que estimulem as discussões acerca da presença *online* de pesquisadores de acordo com as suas particularidades geográficas, evidenciado pela tabela “País x Plataforma”.

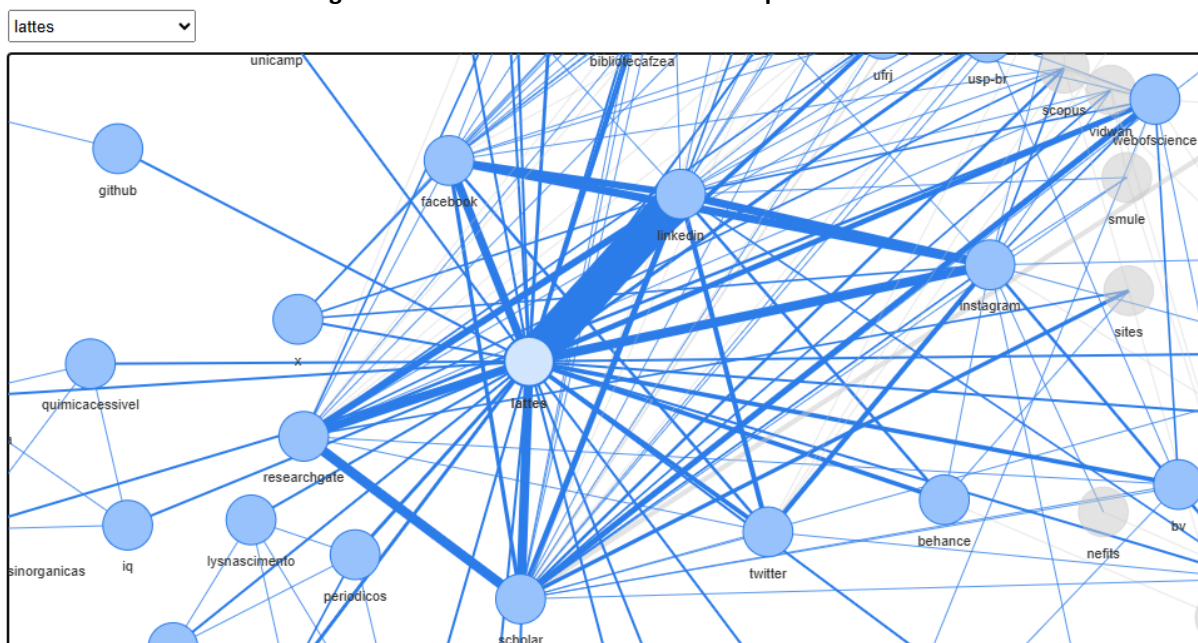
Países como Espanha e Estados Unidos apresentam resultados que indicam uma preferência equilibrada entre mídias sociais acadêmicas (*ResearchGate*, *Google Scholar*) e

mídias sociais gerais (*Instagram*, *X*). Barata *et al.* (2018) menciona que no Brasil os pesquisadores utilizam predominantemente o *Currículo Lattes* como plataforma de identidade acadêmica, ao passo que o uso de redes sociais como *Facebook*, *X* e *ResearchGate* varia conforme as áreas e intenções de visibilidade. Dessa forma a localização influencia a escolha das plataformas, muitas vezes impulsionada por políticas institucionais e culturais específicas do país.

Referente aos dados de frequência de menção das plataformas, destaca-se que foram utilizadas as informações coletadas a partir do campo “Domínio (Plataforma)”, assim pode-se garantir a maior precisão da contagem. As plataformas mais frequentes foram o *Currículo Lattes* (93), *Linkedin* (53), *Google Scholar* (27), *ResearchGate* (24) e *Instagram* (22). Conforme outros estudos, esses resultados destacam a importância das mídias sociais acadêmicas (*ResearchGate*, *Google Scholar*) para aumentar a visibilidade e das mídias sociais gerais (*Linkedin*, *Instagram*) para ampliar o engajamento dos pesquisadores com outros públicos mais amplos (Araújo, 2014; Orduña-Malea, Martín-Martín e Delgado-López-Cózar, 2016).

O campo “*Other IDs*”, onde são inseridas as informações de perfis em outras plataformas que disponibilizam identificadores únicos, apresentou informações em apenas 91 perfis. Destacam-se as plataformas *ResearcherID* (53) e *Scopus Author ID* (38), além das plataformas *Ciência ID* (9), *Sciprofiles* (5) e *Loop profile* (4). Os identificadores *Researcher ID* e *Scopus Author ID* estão vinculados às bases de dados *Web of Science* e *Scopus* respectivamente. Portanto, a existência dessa informação depende da indexação da produção do pesquisador nas referidas bases de dados, assim, observam-se indícios da baixa representação da produção acadêmica destes pesquisadores na *Scopus* e na *Web of Science*. A presença desses identificadores adicionais reforça a ideia apresentada por Costas, Corona e Robinson-García (2024) sobre o papel agregador da ORCID ao vincular diferentes perfis acadêmicos em um único ambiente. Desse modo, a funcionalidade de apresentar tais informações também se mostra igualmente útil para análises de presença *online*. Por fim, apresenta-se a rede de coocorrência das plataformas cadastradas pelos pesquisadores. A Figura 2 apresenta a rede onde o nó do *Currículo Lattes* (lattes) está selecionado.

Figura 2 – Rede de coocorrência entre as plataformas.



Fonte: Elaboração própria.

Analisando a rede acima observa-se que muitos pesquisadores deste universo utilizam principalmente o *Currículo Lattes* como plataforma de presença *online*, seguido por outras plataformas como *ResearchGate*, *Google Scholar* e *Linkedin*. As combinações mais frequentes incluem Lattes com *Instagram* e *Linkedin*; *ResearchGate* com *Google Scholar*; *Facebook* com *Instagram*. A rede de coocorrência confirma a relevância das plataformas destacadas anteriormente e apresenta a possibilidade de analisar visualmente a presença *online* dos pesquisadores.

A partir dessa análise visual observam-se evidências de que o uso simultâneo dessas plataformas pode indicar uma tentativa de maximizar o impacto acadêmico e social. A diversificação da presença em diferentes plataformas é um componente essencial da chamada Ciência 2.0, onde visibilidade, acessibilidade e engajamento com públicos diversos se tornam objetivos relevantes (Araújo, 2014). O uso combinado de plataformas acadêmicas (como ORCID, *ResearchGate* e *Google Scholar*) e redes sociais gerais (como X, *Facebook*, *Linkedin*) não é aleatório, mas sim uma estratégia deliberada de muitos pesquisadores para aumentar sua visibilidade, alcançar diferentes públicos e potencializar seu impacto, tanto na esfera acadêmica quanto na esfera pública e social (Araújo, 2014; Costas; Corona; Robinson-García, 2024).

A análise do grafo pode ser combinada com as informações apresentadas na coluna “Domínio (Plataforma)” na tabela de dados sumarizados para obter uma imagem mais ampla da coocorrência das menções. Observou-se que *Linkedin* e Lattes foram mencionados simultaneamente 36 vezes, *Instagram* e *Linkedin* 6 vezes e *Google Scholar* e Web of Science 5 vezes. Esses resultados sugerem que existe uma forte tendência dos pesquisadores em utilizar conjuntamente uma plataforma acadêmica (Lattes) e uma mídia social profissional (*Linkedin*). Barata *et al.* (2018) corrobora ao mencionar que as redes sociais profissionais são empregadas para fins distintos das redes acadêmicas, e, nesse contexto, o *Linkedin* aparece como uma ferramenta relevante para a construção da reputação profissional, colaboração e divulgação de conquistas e atividades acadêmicas em ambientes mais amplos.

4 CONSIDERAÇÕES FINAIS

A partir da prova de conceito pode-se demonstrar que o algoritmo desenvolvido é eficaz na identificação automática da presença *online* de pesquisadores registrados na ORCID. A análise dos dados permitiu a caracterização da presença digital dos pesquisadores, identificando padrões geográficos, preferências por plataformas e a integração entre diferentes identificadores acadêmicos. Dessa forma, o estudo responde diretamente para a questão norteadora de como identificar automaticamente a presença *online* de pesquisadores em diferentes ambientes web.

Os resultados reforçam a importância da gestão da presença *online* dos pesquisadores. O estudo evidenciou que a presença *online* dos pesquisadores varia de acordo com o país, sendo o *Currículo Lattes* predominante no universo analisado, enquanto *ResearchGate* e *Google Scholar* têm maior relevância em outros países. A análise da rede de coocorrência evidenciou padrões de uso combinados entre mídias sociais acadêmicas e de uso pessoal ou profissional, demonstrando a relevância da diversificação da presença *online* para ampliar a visibilidade dos pesquisadores.

Entre as limitações do estudo, destaca-se a dependência da atualização das informações nos perfis da ORCID, uma vez que a metodologia utilizada se baseia nos dados fornecidos voluntariamente pelos pesquisadores. Outra possível limitação refere-se à própria definição de presença *online*, vale ressaltar que a simples menção de perfis de outras

plataformas na ORCID não garante que essa presença tenha qualidade, ou seja, são necessários outros estudos que possam verificar se os perfis indicados estão atualizados e se os dados são válidos. Assim, definiu-se que, para o objetivo proposto, a menção de uma url ou ID nos campos indicados na ORCID representa um indício de presença *online* daquele pesquisador.

Enquanto trabalhos futuros, pretende-se realizar a validação dos dados a fim de analisar a qualidade da presença *online* desses pesquisadores verificando se os perfis estão ativos e atualizados. Além disso, pretende-se desenvolver uma interface gráfica que permita o uso do algoritmo sem a necessidade de interação direta com o *script* do algoritmo, tornando a ferramenta mais acessível. Os resultados reforçam a relevância da ORCID como um agregador de perfis acadêmicos e demonstram o potencial de abordagens automatizadas para mapear a presença *online* de pesquisadores. Por fim, ao fornecer um método estruturado para coletar e analisar esses dados, este estudo contribui para a compreensão das dinâmicas de visibilidade acadêmica na web e para o aprimoramento de estratégias voltadas à presença *online* de pesquisadores.

REFERÊNCIAS

- ARAÚJO, R. F. Ciência 2.0 e a presença *online* de pesquisadores: visibilidade e impacto. **Ciência da Informação em Revista**, Maceió, v. 1, n. 3, p. 32-40, 2014. DOI; <http://doi.org/10.28998/cirev.%y132-40>. Disponível em: <https://www.seer.ufal.br/index.php/cir/article/view/1608>. Acesso em: 28 fev. 2025.
- ARAÚJO, R. F.; PEDRI, P. Cientistas da informação no Twitter: presença, visibilidade e autoridade social. **P2P e Inovação**, Rio de Janeiro, v. 10, n. 1, p. 371-385, 2023. DOI; <http://doi.org/10.21728/p2p.2023v10n1.p371-385>. Disponível em: <https://cip.brapci.inf.br/download/228365>. Acesso em: 9 abr. 2025.
- ARROYO-MACHADO, W. *et al.* Exploring ORCID adoption and metadata presence in Spain's research landscape. **Scientometrics**, [S. l.], v. 130, n. 1, p. 123-145, 2025. DOI: <http://doi.org/10.1007/s11192-025-05248-8>. Disponível em: <https://link.springer.com/article/10.1007/s11192-025-05248-8>. Acesso em: 7 mar. 2025.
- CASTANHA, R. G.; FRANCO, F. Conjunto de dados utilizados no estudo "algoritmo para análise de presença online de pesquisadores da base ORCID". **Zenodo**. Version v1. 2025. DOI: <https://doi.org/10.5281/zenodo.15001663>. Acesso em: 11 maio 2025.

BARATA, G. *et al.* O uso de mídias sociais por acadêmicos brasileiros. *In: ENCONTRO BRASILEIRO DE BIBLIOMETRIA E CIENTOMETRIA*, 6., 2018, Rio de Janeiro. **Anais [...]**. Rio de Janeiro: EBBC, 2018. p. 209-217. DOI: <http://doi.org/10.22477.047989>. Disponível em: <https://ebbc.inf.br/ojs/index.php/ebbc/article/view/715>. Acesso em: 7 mar. 2025.

BIK, Holly M.; GOLDSTEIN, Miriam C. An introduction to social media for scientists. **PLOS Biology**, [S. l.], v. 11, n. 4, e1001535, 2013. Disponível em: <https://doi.org/10.1371/journal.pbio.1001535>. Acesso em: 9 abr. 2025.

CARNEIRO, A. M.; PINTO, D. M.; GOMES, J. G. P. M. ORCID como fonte de dados para análise da carreira acadêmica no Brasil: oportunidades e desafios. *In: ENCONTRO BRASILEIRO DE BIBLIOMETRIA E CIENTOMETRIA*, 9., 2024, Brasília, DF. **Anais [...]**. Brasília, DF: EBBC, 2024. Disponível em: <https://doi.org/10.22477/ix.ebbc.360>. Acesso em: 6 mar. 2025.

COSTAS, R.; CORONA, C.; ROBINSON-GARCÍA, N. Could ORCID play a key role in meta-research? discussing new analytical possibilities to study the dynamics of science and scientists. *In: OANCEA, A.; DERRICK, G. E.; NUSEIBEH, N.; XU, X. (org.). Handbook of meta-research*. Cheltenham: Edward Elgar Publishing, 2024. cap. 18, p.215-232.

ORCID. **ORCID Statistics**. 2025. Disponível em: <https://info.ORCID.org/pt/resources/ORCID-statistics/>. Acesso em: 6 mar. 2025.

ORDUÑA-MALEA, E.; MARTÍN-MARTÍN, A.; DELGADO-LÓPEZ-CÓZAR, E. The next bibliometrics: ALMetrics (author level metrics) and the multiple faces of author impact. **Profesional de la información**, [S. l.], v. 25, n. 3, p. 485-496, 2016. Disponível em: <https://doi.org/10.3145/epi.2016.may.18>. Acesso em: 28 fev. 2025.

SIXTO-COSTOYA, A. *et al.* Exploring the relevance of ORCID as a source of study of data sharing activities at the individual-level: a methodological discussion. **Scientometrics**, [S. l.], v. 126, n. 8, p. 7149-7165, 2021. Disponível em: <https://arxiv.org/abs/2105.11825>. Acesso em: 8 mar. 2025.

WANG, T. *et al.* Is ORCID a reliable source for CV analysis? exploring the data availability of ORCID academic profiles. **Scientometrics**, [S. l.], v. 129, n. 3, p. 1637-1662, 2024. Disponível em: <https://doi.org/10.1007/s11192-024-04944-1>. Acesso em: 28 fev. 2025.



4.0 Internacional

Esta licença permite compartilhamento, remixe, adaptação e criação a partir do trabalho, mesmo para fins comerciais, desde que sejam atribuídos créditos ao(s) autor(es). Conteúdos elaborados por terceiros, citados e referenciados nesta obra não são cobertos pela licença.

Como citar este trabalho:

CASTANHA, Rafael Gutierrez; FRANCO, Francielle. Algoritmo para análise de presença *online* de pesquisadores da base ORCID. *In: WORKSHOP DE INFORMAÇÃO DADOS E TECNOLOGIA*, 8., 2025, Marília, SP. **Anais [...]**. Marília, SP: Universidade de Marília, 2025. DOI: <https://doi.org/10.22477/viii.widat.265>.



WIDaT 2025

VIII WORKSHOP DE INFORMAÇÃO,
DADOS E TECNOLOGIA

MARÍLIA • SP

Qualidade de dados: análise focada na categoria representacional

Data quality: analysis focused on the representational category

Débora Kaori Yoshitane¹



<http://lattes.cnpq.br/2099435419438744>



<https://orcid.org/0009-0005-4214-3602>

Ananda Fernanda de Jesus²



<http://lattes.cnpq.br/8553935355036462>



<https://orcid.org/0000-0001-7873-6040>

José Eduardo Santarem Segundo³



<http://lattes.cnpq.br/5562746387565465>



<https://orcid.org/0000-0003-3360-7872>

Resumo

A qualidade de dados pode ser dividida em quatro categorias: intrínseca, contextual, representacional e acessibilidade. A qualidade representacional é definida a partir de quão bem estruturados estão os conjuntos de dados, avaliando principalmente aspectos relacionados ao formato e significado desses dados. Esta pesquisa tem como objetivo verificar o estado da arte das discussões a respeito da categoria representacional em pesquisas de qualidade de dados, buscando identificar como o assunto tem sido discutido pela literatura e potenciais temas a serem explorados. Como metodologia, utilizou-se a análise exploratória baseada em um protocolo de pesquisa. Os resultados mostram a preocupação de diversos setores com a qualidade de dados, apontando a qualidade representacional como um fator importante na percepção dos consumidores de dados para tomadas de decisão. Conclui-se que a qualidade representacional de dados ainda é um assunto pouco explorado isoladamente. Conclui-se ainda que a definição do conceito de qualidade de dados representacional, suas dimensões, critérios e métricas ainda podem ser estudados de maneira mais aprofundada.

Palavras-chave: qualidade de dados; qualidade representacional; análise exploratória.

¹ Graduanda em Biblioteconomia e Ciência da Informação, Universidade de São Paulo (USP), Ribeirão Preto, São Paulo, Brasil. Bolsista do Programa Unificado de Bolsas (PUB-USP). deborayoshitane@usp.br.

² Doutoranda em Ciência da Informação, Universidade Estadual Paulista (UNESP), Marília, São Paulo, Brasil. Bolsista da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP). af.jesus@unesp.br.

³ Livre Docente em Informação e Tecnologia pela Universidade de São Paulo (USP), Ribeirão Preto, São Paulo, Brasil. Professor Associado, Universidade de São Paulo (USP). santarem@usp.br.

Abstract

Data quality can be divided into four categories: intrinsic, contextual, representational, and accessibility. Representational quality is defined based on how well structured the data sets are, mainly evaluating aspects related to the format and meaning of these data. This research aims to verify the state of the art of discussions regarding the representational dimension in data quality research, seeking to identify how the subject has been discussed in the literature and potential themes to be explored. The methodology used was an exploratory analysis based on a research protocol. The results show the concern of several sectors with data quality, pointing to representational quality as an important factor in the perception of data consumers for decision-making. It is concluded that the representational quality of data is still a subject little explored in isolation, and that the definition of the concept of representational quality, its dimensions, criteria, and metrics can still be further deepened and studied.

Keywords: data quality; representational quality; exploratory analysis.

1 INTRODUÇÃO

Tendo em vista as consequências de problemas de qualidade de dados, como impactos econômicos e sociais, afetando principalmente a tomada de decisão de organizações, a preocupação com a qualidade é uma questão bastante discutida atualmente, especialmente relacionada com a produção e com o consumo de dados em grande quantidade.

A definição adotada com maior frequência na literatura atual de qualidade de dados é associada ao conceito de *fitness for use*, em que a qualidade é entendida a partir de uma abordagem contextual, definindo a qualidade de dados como sinónimo de dados considerados adequados para uso pelos consumidores (Juran, 1988; Zaveri *et al.*, 2012; Langer *et al.*, 2018). Essa definição, centrada no consumidor, também é abordada por Redman (2013), que afirma que uma coleção de dados é de alta qualidade desde que atenda as necessidades do seu consumidor.

Embora os aspectos contextuais tenham um forte impacto, a qualidade de dados pode ser entendida, estudada e avaliada a partir de outras perspectivas, baseadas nas categorias de qualidade estabelecidas por Wang e Strong (1996).

A estrutura da qualidade de dados enquanto uma área de investigação é fortemente influenciada pelas pesquisas de Wang e Strong (1996). Os autores buscaram discutir e definir dimensões, critérios e métricas para o processo de avaliação de qualidade de dados, organizando-os em diferentes categorias que refletem as distintas perspectivas da qualidade dos dados, levando em consideração não apenas a sua acurácia, mas também as necessidades dos usuários. Suas contribuições servem como

base para diversos estudos mais recentes, fato evidenciado pelo grande número de citações de seus trabalhos (Sadiq, 2013, p. 7).

Sendo assim, Wang e Strong (1996) definiram quatro categorias de qualidade de dados segundo aspectos apontados pelos consumidores: intrínseca, contextual, representacional e acessibilidade.

A qualidade representacional, objeto de estudo desta pesquisa, diz respeito ao formato e significado dos dados:

Esses dois aspectos sugerem que, para que os consumidores de dados concluam que os dados estão bem representados, eles devem ser não apenas concisos e consistentemente representados, mas também interpretáveis e fáceis de entender (Wang; Strong, 1996, p. 21).

Embora a qualidade representacional tenha se consolidado, a partir da pesquisa de Wang e Strong (1996), como uma das quatro categorias de qualidade de dados e seja frequentemente mencionada em estudos na área da qualidade de dados, as pesquisas geralmente se concentram nos aspectos intrínsecos e contextuais da qualidade de dados, sendo as dimensões representacional e acessibilidade significativamente menos exploradas pela literatura. Nesse sentido, a elaboração do presente estudo partiu do questionamento: Como é discutida na literatura a dimensão representacional da qualidade de dados?

O objetivo desta pesquisa é verificar, por meio de uma análise exploratória da literatura, com adoção de protocolo de pesquisa, como tem sido discutida a categoria representacional, buscando identificar os temas mais abordados, bem como temas ainda para serem explorados. A próxima seção apresenta os procedimentos metodológicos adotados para a condução da pesquisa.

2 PROCEDIMENTOS METODOLÓGICOS

Considerando seus objetivos e procedimentos metodológicos, esta pesquisa se caracteriza por sua natureza exploratória, tendo como finalidade reunir informações preliminares e aprofundar a compreensão teórica, estabelecendo assim uma base para investigações futuras (Prodanov; Freitas, 2013). Quanto a seus resultados, a pesquisa se caracteriza por seu caráter qualitativo e descritivo.

Para alcançar o objetivo proposto, foi realizada uma análise exploratória da literatura científica, baseada na adoção de um protocolo de pesquisa. Esse protocolo foi elaborado visando orientar a busca, seleção e análise de um *corpus* teórico que permita uma maior compreensão a respeito de como o assunto tem sido elaborado. A análise exploratória é considerada como uma etapa inicial e indispensável para outros estudos mais aprofundados da literatura, podendo seu protocolo ser considerado como protocolo piloto para Revisões Sistemáticas da Literatura e Mapeamentos Sistemáticos da Literatura. O Quadro 1 apresenta o protocolo elaborado para a pesquisa:

Quadro 1 - Protocolo de pesquisa

Protocolo de busca	
Pergunta de pesquisa (principal)	Como é discutida na literatura a dimensão representacional da qualidade de dados?
Objetivos	Identificar o estado da arte das discussões a respeito da dimensão representacional em pesquisas de qualidade de dados
Bases de dados consultada	Web of Science, Taylor & Francis, LISA, LISTA
Estratégia de busca	“Data quality” AND “Representational”
Período abrangido	Sem recorte temporal
Idioma	Inglês
Tipo documental	Artigos em periódicos
CrITÉRIOS de inclusão	(I) Discute a dimensão representacional da qualidade de dados (I) Foco na dimensão representacional da qualidade de dados (I) Apresenta uma pesquisa aplicada de qualidade de dados com base uma dimensão representacional
CrITÉRIOS de exclusão	(E) Não está nos idiomas estabelecidos para a pesquisa (E) Não está no formato estabelecido para a pesquisa (E) Apenas menciona a dimensão representacional (E) Não aplica ou não discute o tema da pesquisa (E) Não é possível ter acesso ao documento completo
Formulário de extração	1) Objetivo principal do estudo; 2) Forma como a temática foi abordada; 3) Definições
Data da coleta	Outubro de 2024

Fonte: Elaboração própria.

Com a finalidade de ampliar o *corpus* teórico construído foi adotada ainda a técnica do *snowballing*, que permitiu a identificação e inclusão de referências dos artigos pertinentes ao tema. Apresentados os procedimentos metodológicos adotados, a próxima seção apresenta os resultados obtidos.

3 RESULTADOS E DISCUSSÃO

Com base no protocolo apresentado, 320 artigos foram recuperados. Destes, 294 foram recusados, 16 foram aceitos e 10 eram artigos duplicados. Com o *snowballing*, mais 9 artigos foram incluídos, totalizando 25 artigos para análise (Gráfico 1).



Fonte: Elaboração própria.

Os artigos, então, foram lidos e organizados pelo ano de publicação. O Quadro 2 apresenta o objetivo dos estudos selecionados, seus autores e sua fonte, destacando se foram recuperados pela estratégia de busca ou pelo *snowballing*. Os destaques mostram os trabalhos focados apenas na dimensão representacional.

Quadro 2 - Objetivos dos estudos selecionados

Autores	Objetivo	Fonte
Wang e Strong (1996)	Apresentar um framework com aspectos da qualidade de dados considerados importantes pelos consumidores de dados	Estratégia de busca
Strong, Lee e Wang (1997)	Apresentar uma análise de projetos de qualidade de dados e identificação dos padrões mais comuns de problemas de qualidade	<i>Snowballing</i>
Klein (2001)	Comparar as percepções do usuário em relação à qualidade de dados encontrada na internet e em fontes tradicionais de texto	Estratégia de busca
Calero, Caro e Piattini (2008)	Apresentar uma ferramenta de avaliação de qualidade de dados para portais da web	<i>Snowballing</i>

Caro <i>et al.</i> (2008)	Identificar atributos relevantes para avaliação da qualidade de dados de portais da web	<i>Snowballing</i>
Blake e Mangiameli (2009)	Apresentar métricas para consistência representacional de dados estruturados e não estruturados	<i>Snowballing</i>
Moraga <i>et al.</i> (2009)	Apresentar o desenvolvimento de um modelo de qualidade para portais da web	<i>Snowballing</i>
Salomone, Hyland e Murphy (2011)	Analisar atributos da qualidade de dados em relação a diferentes funções de pessoas que trabalham com dados	<i>Snowballing</i>
Zaveri <i>et al.</i> (2015)	Realizar uma revisão sistemática sobre métodos de avaliação de qualidade de linked data	<i>Snowballing</i>
Heo (2013)	Testar um modelo de avaliação de portais de patrimônio cultural e verificar sua relação com a satisfação do usuário	Estratégia de busca
Leite <i>et al.</i> (2014)	Propor um modelo de avaliação de qualidade de dados em sites de unidades de saúde	Estratégia de busca
Yi, Hua e Sadiq (2016)	Propor um framework para lidar com a inconsistência representacional de dados	Estratégia de busca
Liu <i>et al.</i> (2017)	Investigar se o gênero influencia na avaliação da qualidade de informação em comunidades virtuais com uma perspectiva de expectativa-percepção	Estratégia de busca
Juddoo <i>et al.</i> (2018)	Discutir as dimensões de qualidade de dados no contexto de big data para a indústria da saúde	<i>Snowballing</i>
Langer <i>et al.</i> (2018)	Apresentar uma ferramenta de avaliação de qualidade, o SemQuire	Estratégia de busca
Shang, Weng e Hripcsak (2018)	Apresentar um framework para ajudar na escolha de bases de dados clínicas adequadas para conduzir estudos observacionais	Estratégia de busca
Wahyudi, Farhani e Janssen (2018)	Comparar características de big data com as dimensões da qualidade de dados de organizações de serviços financeiros	<i>Snowballing</i>
Ghasemaghaei e Calic (2019)	Explorar o impacto do uso de big data na qualidade de dados para analisar a qualidade da tomada de decisão de empresas	Estratégia de busca
Kwiatkowska e Pouw (2019)	Propor um framework a partir de uma perspectiva semiótica para qualidade de dados	Estratégia de busca
Mulgund <i>et al.</i> (2020)	Realizar uma revisão sistemática para identificar problemas de qualidade de dados em sites de avaliação de médicos	Estratégia de busca
Jabar <i>et al.</i> (2022)	Investigar e explorar as dimensões de qualidade de dados aplicadas à análise de big data no setor público da Malásia	Estratégia de busca

Zhang e Trace (2022)	Identificar quais dimensões da qualidade de dados os consumidores consideram importantes para avaliar a qualidade de informação dos seus próprios dados de saúde	Estratégia de busca
Matos, Jorge e Sá (2023)	Apresentar uma ferramenta para medir a qualidade das informações de planejamento financeiro estratégico de municípios de Portugal	Estratégia de busca
Priestley, O'Donnell e Simperl (2023)	Apresentar requisitos de qualidade de dados importantes para sistemas de machine learning em cada estágio de desenvolvimento	Estratégia de busca
Timmerman <i>et al.</i> (2023)	Apresentar uma abordagem baseada em custos para analisar e medir a qualidade de dados a partir da consistência representacional	Estratégia de busca

Fonte: Elaboração própria.

A maioria dos artigos recuperados é da área da Ciência da Computação, mas a preocupação com a qualidade de dados não se limita em apenas uma área. Dentre os temas encontrados relacionados à qualidade de dados, destaca-se: a qualidade de dados no contexto do *big data* e do *linked data*; a qualidade de dados médicos e da saúde; a qualidade de dados em sistemas de informação; a qualidade de dados nos setores públicos, financeiros e industriais; e a qualidade de dados da web.

Analisando os objetivos dos estudos, é possível notar que a maioria aborda a qualidade representacional juntamente com as outras categorias, visando discutir diferentes aspectos da qualidade de dados. Mesmo em diversos contextos, grande parte dos trabalhos analisados diretamente baseia suas definições de qualidade no artigo de Wang e Strong (1996), confirmando a influência que esses autores ainda exercem sobre o assunto. De acordo com as dimensões da qualidade representacional (interpretabilidade, facilidade de compreensão, consistência representacional e representação concisa), os autores afirmam que “problemas relacionados ao significado e ao formato surgem na pesquisa de sistemas de banco de dados, na qual o formato é abordado como parte da sintaxe, e o significado, como parte da reconciliação semântica” (Wang; Strong, 1996, p. 21).

Dentre todos os documentos que compõem o *corpus* da pesquisa, apenas três têm como foco principal a qualidade representacional de dados (títulos em destaque no Quadro 2). Ainda assim, estes trabalhos analisam somente a dimensão da consistência representacional.

Yi, Hua e Sadiq (2016) analisaram a inconsistência representacional, apontando o desafio de manter a consistência dos dados devido aos seus diferentes formatos e estruturas, além de serem gerados e representados de forma diferente em diversos países, culturas e empresas.

Quando estes dados distintos são integrados, a inconsistência dos dados torna-se evidente. Por esse motivo, este problema está atraindo cada vez mais a atenção da indústria e da academia, com os dados se tornando cada vez mais massivos e heterogêneos. As pesquisas atuais sobre inconsistência de representação de dados geralmente se concentram em um domínio específico, como nome, endereço etc. (Yi; Hua; Sadiq, 2016, p. 395).

Para solucionar esse problema, as autoras sugerem um framework de três módulos (design, detecção e uniformização de padrões) que gera possíveis padrões a partir de dados apresentados em diversos domínios.

Timmerman *et al.* (2023) aplicaram um experimento com 218 voluntários visando investigar o impacto que a representação consistente e a completude dos dados têm na qualidade. Os resultados mostraram um aumento na usabilidade dos dados ao melhorar sua qualidade por meio da correção de problemas de completude e consistência representacional.

A análise baseada em custos foi implementada na prática por meio de um experimento para a análise de um caso de uso que consistiu em três conjuntos de dados com problemas de integridade e consistência de representação. O experimento confirmou (e quantificou) o aumento esperado da utilidade dos dados devido às melhorias na qualidade dos dados. (Timmerman *et al.*, 2023, p. 10).

De um modo geral, a dimensão representacional é apresentada como um fator essencial para os consumidores de dados: a interpretabilidade, a facilidade de compreensão e a representação concisa e consistente dos dados podem influenciar diretamente nas tomadas de decisão. Uma má qualidade representacional de dados pode resultar em falhas na comunicação, decisões prejudiciais e impactos negativos nos processos de trabalho. Para evitar isso, os estudos propõem definições de padrões adequados ao seu contexto e a documentação das padronizações.

4 CONSIDERAÇÕES FINAIS

Por meio da condução de uma análise exploratória da literatura e da adoção da técnica de *snowballing*, esta pesquisa teve como objetivo identificar de que modo a literatura discute a qualidade representacional de dados. Outro objetivo deste estudo foi verificar a relevância da qualidade representacional, visando compreender seu impacto nos processos de avaliação de qualidade de dados.

Dessa forma, a qualidade representacional de dados mostrou-se como um aspecto importante na percepção dos consumidores de dados, exigindo que estes sejam fáceis de entender e de interpretar, além de serem representados de maneira concisa e consistente. Problemas na qualidade representacional podem trazer prejuízos aos envolvidos, tendo em vista que uma representação falha impede a utilização eficiente dos dados.

A diversidade de contextos dos estudos recuperados pela estratégia de busca demonstra o interesse amplo pela qualidade de dados, principalmente no contexto do big data e dados no contexto da web, no qual o grande volume e a diversidade de fontes dos dados apresentam-se como principais desafios a serem resolvidos na questão da qualidade representacional.

Apesar de sua relevância e impacto na percepção da qualidade em conjuntos de dados, conclui-se que a categoria representacional ainda é pouco explorada pelos pesquisadores, especialmente pesquisas focadas nesta categoria que busquem apresentar definições claras a respeito da qualidade representacional, identificar dimensões e métricas para a sua avaliação, elaborar e testar ferramentas que auxiliem na avaliação deste aspecto de qualidade.

Nesse sentido, a próxima etapa será realizar ajustes no protocolo de pesquisa, visando aplicar uma Revisão Sistemática da Literatura. Ressalta-se a necessidade de um aprofundamento em relação à definição do termo qualidade representacional, visando verificar se esse termo tem sido adotado com significado consistente pelos autores, buscando ainda fornecer maior clareza terminológica a respeito da aplicação do termo. Como estudos futuros, pretende-se explorar mais detalhadamente as dimensões, critérios e métricas relacionados com a qualidade representacional.

Outro aspecto que demanda maior aprofundamento é a potencial relação entre os aspectos representacionais da qualidade de dados e a Ciência da Informação (CI). Ressalta-se que em ambas as áreas o conceito de representação pode assumir diferentes acepções, o que pode ter implicações sobre como essa temática pode ser entendida e discutida no âmbito da CI.

Os resultados desta análise exploratória apontam que a categoria representacional é essencial para uma avaliação completa de qualidade de dados, apesar de ser pouco explorada pela ciência. Ao focar na qualidade representacional, esta pesquisa recupera um assunto ainda a ser desenvolvido em profundidade, destacando suas possíveis definições, dimensões, critérios e soluções para a qualidade representacional dos dados.

REFERÊNCIAS

- BLAKE, Roger H.; MANGIAMELI, Paul. Evaluating the semantic and representational consistency of interconnected structured and unstructured data. **AMCIS 2009 Proceedings**, n. 126, 2009. Disponível em: <https://aisel.aisnet.org/amcis2009/126>. Acesso em: 12 maio 2025.
- CALERO, Coral; CARO, Angélica; PIATTINI, Mario. An applicable data quality model for web portal data consumers. **World Wide Web**, v. 11, p. 465-484, 2008. Disponível em: <https://link.springer.com/article/10.1007/s11280-008-0048-y>. Acesso em: 12 maio 2025.
- CARO, Angélica *et al.* A proposal for a set of attributes relevant for web portal data quality. **Software Quality Journal**, v. 16, p. 513-542, 2008. Disponível em: <https://link.springer.com/article/10.1007/s11219-008-9046-7>. Acesso em: 12 maio 2025.
- GHASEMAGHAEI, Maryam; CALIC, Goran. Can big data improve firm decision quality? The role of data quality and data diagnosticity. **Decision Support Systems**, v. 120, p. 38-49, 2019. Disponível em: <https://dl.acm.org/doi/abs/10.1016/j.dss.2019.03.008>. Acesso em: 12 maio 2025.
- HEO, Misook. User satisfaction with portals: testing for factorial validity and invariance across age groups. **Online Information Review**, v. 37, n. 5, p. 804-820, 2013. Disponível em: <https://www.emerald.com/insight/content/doi/10.1108/oir-06-2012-0099/full/html>. Acesso em: 12 maio 2025.

JABAR, Zam Zarina Abdul *et al.* Establishment of big data analytics application model for Malaysian public sector: an expert validation. **EDPACS**, v. 65, n. 4, p. 1-20, 2022.

Disponível em:

<https://www.tandfonline.com/doi/full/10.1080/07366981.2021.1958736>. Acesso em: 12 maio 2025.

JUDDOO, Suraj *et al.* Data governance in the health industry: investigating data quality dimensions within a big data context. **Applied System Innovation**, v. 1, n. 4, p. 43, 2018.

Disponível em: <https://www.mdpi.com/2571-5577/1/4/43>. Acesso em: 12 maio 2025.

JURAN, Joseph Moses. **Quality control handbook**. New York: Mcgraw-Hill, 1988.

KLEIN, Barbara D. User perceptions of data quality: Internet and traditional text sources. **Journal of computer information systems**, v. 41, n. 4, p. 9-15, 2001. Disponível em:

<https://www.tandfonline.com/doi/abs/10.1080/08874417.2001.11647016>. Acesso em: 12 maio 2025.

KWIATKOWSKA, Mila; POUW, Frank. Data quality in secondary data analysis: a case study of ecological data using a semiotic-based approach. *In*: INTERNATIONAL CONFERENCE ON DATA SCIENCE, TECHNOLOGY AND APPLICATIONS, 8., 2019. Prague.

Proceedings [...]. Prague: [S. l.], 2019. v. 1, p. 377-384. Disponível em:

<https://www.scitepress.org/Link.aspx?doi=10.5220/0007978403770384>. Acesso em: 12 maio 2025.

LANGER, André *et al.* SemQuire: assessing the data quality of linked open data sources based on DQV. **Current Trends in Web Engineering**, v. 11153, p. 163-175, 2018.

Disponível em: https://link.springer.com/chapter/10.1007/978-3-030-03056-8_14.

Acesso em: 12 maio 2025.

LEITE, Patrícia *et al.* Towards a model for the measurement of data quality in websites.

New Review of Hypermedia and Multimedia, v. 20, n. 4, p. 301-316, 2014. Disponível

em: <https://www.tandfonline.com/doi/full/10.1080/13614568.2014.968638>. Acesso em: 12 maio 2025.

LIU, Yuewen *et al.* Gender differences in information quality of virtual communities: a study from an expectation-perception perspective. **Personality and individual differences**, v. 104, p. 224-229, 2017. Disponível em:

<https://www.sciencedirect.com/science/article/pii/S0191886916309084>. Acesso em: 12 maio 2025.

MATOS, Sandra; JORGE, Susana; SÁ, Patricia Moura e. Measuring the quality of the strategic financial planning information (Q-FPI) in the local government. **International Journal of Public Administration**, v. 46, n. 10, p. 671-688, 2023. Disponível em:

<https://www.tandfonline.com/doi/full/10.1080/01900692.2021.2011315>. Acesso em: 12 maio 2025.

- MORAGA, Carmen *et al.* SQuaRE-aligned data quality model for web portals. *In*: INTERNATIONAL CONFERENCE ON QUALITY SOFTWARE, 9., 2009. Jeju, Korea (South). **Proceedings [...]**. Jeju, Korea (South): IEEE, 2009. p. 117-122. Disponível em: <https://ieeexplore.ieee.org/document/5381502>. Acesso em: 12 maio 2025.
- MULGUND, Pavankumar *et al.* Data quality issues with physician-rating websites: systematic review. **Journal of Medical Internet Research**, v. 22, n. 9, p. e15916, 2020. Disponível em: <https://www.jmir.org/2020/9/e15916/>. Acesso em: 12 maio 2025.
- PRIESTLEY, Maria; O'DONNELL, Fionntán; SIMPERL, Elena. A survey of data quality requirements that matter in ML development pipelines. **ACM Journal of Data and Information Quality**, v. 15, n. 2, p. 1-39, 2023. Disponível em: <https://dl.acm.org/doi/10.1145/3592616>. Acesso em: 12 maio 2025.
- PRODANOV, Cleber Cristiano; FREITAS, Ernani Cesar de. **Metodologia do trabalho científico: métodos e técnicas da pesquisa e do trabalho acadêmico**. 2. ed. Novo Hamburgo: Editora Feevale, 2013.
- REDMAN, Thomas C. Data quality management past, present, and future: towards a management system for data. *In*: SADIQ, Shazia (ed.). **Handbook of data quality: research and practice**. Berlin: Springer Berlin Heidelberg, 2013. p. 15-40. Disponível em: https://link.springer.com/chapter/10.1007/978-3-642-36257-6_2. Acesso em: 12 maio 2025.
- SADIQ, Shazia. Prologue: research and practice in data quality management. *In*: SADIQ, Shazia (ed.). **Handbook of data quality: research and practice**. Berlin: Springer Berlin Heidelberg, 2013. p. 1-11. Disponível em: https://link.springer.com/chapter/10.1007/978-3-642-36257-6_1. Acesso em: 12 maio 2025.
- SALOMONE, Sonia; HYLAND, Paul; MURPHY, Glen. Perceptions of data quality dimensions and data roles. *In*: AUSTRALIAN AND NEW ZEALAND ACADEMY OF MANAGEMENT CONFERENCE, 25., 2011, Wellington. **Proceedings [...]**. Wellington: [s.n.], 2011. p. 1-25. Disponível em: <https://eprints.qut.edu.au/48014/>. Acesso em: 12 maio 2025.
- SHANG, Ning; WENG, Chunhua; HRIPCSAK, George. A conceptual framework for evaluating data suitability for observational studies. **Journal of the American Medical Informatics Association**, v. 25, n. 3, p. 248-258, 2018. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/29024976/>. Acesso em: 12 maio 2025.
- STRONG, Diane M.; LEE, Yang W.; WANG, Richard Y. Data quality in context. **Communications of the ACM**, v. 40, n. 5, p. 103-110, 1997. Disponível em: <https://dl.acm.org/doi/10.1145/253769.253804>. Acesso em: 12 maio 2025.
- TIMMERMAN, Yoram *et al.* Cost-based analysis of the impact of data completeness and representational consistency. **Decision Support Systems**, v. 175, p. 114044, 2023.

Disponível em: <https://www.sciencedirect.com/science/article/pii/S0167923623001197>. Acesso em: 12 maio 2025.

WAHYUDI, Agung; FARHANI, Adiska; JANSSEN, Marijn. Relating big data and data quality in financial service organizations. *In: INTERNATIONAL CONFERENCE ON E-BUSINESS, E-SERVICES AND E-SOCIETY*, 17., 2018, Tirana. **Proceedings [...]**. [S. l.]: Springer, 2018. p. 504-519. Disponível em: <https://repository.tudelft.nl/record/uuid:f8bd8560-c503-466e-82d4-069141188061>. Acesso em: 12 maio 2025.

WANG, Richard Y.; STRONG, Diane M. Beyond accuracy: what data quality means to data consumers. **Journal of Management Information Systems**, v. 12, n. 4, p. 5-33, 1996. Disponível em: <http://www.jstor.org/stable/40398176>. Acesso em: 12 maio 2025.

YI, Bingyu; HUA, Wen; SADIQ, Shazia. A pattern-based framework for addressing data representational inconsistency. **Databases Theory and Applications**, v. 9877, p. 395-406, 2016. Disponível em: https://link.springer.com/chapter/10.1007/978-3-319-46922-5_31. Acesso em: 12 maio 2025.

ZAVERI, Amrapali *et al.* Quality assessment for linked open data: a survey. **Semantic Web Journal**, v. 7, n. 1, p. 63-93, 2015. Disponível em: <https://journals.sagepub.com/doi/full/10.3233/SW-150175>. Acesso em: 12 maio 2025.

ZHANG, Yan; TRACE, Ciaran B. The quality of health and wellness self-tracking data: a consumer perspective. **Journal of the Association for Information Science and Technology**, v. 73, n. 6, p. 879-891, 2022. Disponível em: <https://asistdl.onlinelibrary.wiley.com/doi/full/10.1002/asi.24591>. Acesso em: 12 maio 2025.

AGRADECIMENTOS

Agradecemos ao PUB-USP e à FAPESP (Processo 2021/03349-0) pelo financiamento desta pesquisa.



[4.0 Internacional](https://creativecommons.org/licenses/by/4.0/)

Esta licença permite compartilhamento, remixe, adaptação e criação a partir do trabalho, mesmo para fins comerciais, desde que sejam atribuídos créditos ao(s) autor(es). Conteúdos elaborados por terceiros, citados e referenciados nesta obra não são cobertos pela licença.

Como citar este trabalho:

YOSHITANE, Débora Kaori; JESUS, Ananda Fernanda de; SANTAREM SEGUNDO, José Eduardo. Qualidade de dados: análise focada na categoria representacional. *In: WORKSHOP DE INFORMAÇÃO DADOS E TECNOLOGIA*, 8., 2025, Marília, SP. **Anais [...]**. Marília, SP: Universidade de Marília, 2025. DOI: <http://doi.org/10.22477/viii.widat.221>.



WIDaT 2025

VIII WORKSHOP DE INFORMAÇÃO,
DADOS E TECNOLOGIA

MARÍLIA • SP

Reflexões sobre Inteligência Artificial Generativa e os direitos autorais

Reflections on Generative Artificial Intelligence and copyright

Rosilene Paiva Marinho de Sousa¹



<http://lattes.cnpq.br/4465533418771961>



<https://orcid.org/0000-0002-4699-8692>

Natalia Nakano²



<http://lattes.cnpq.br/4612974888327511>



<https://orcid.org/0000-0002-3217-2515>

Milton Shintaku³



<http://lattes.cnpq.br/8605833104600600>



<https://orcid.org/0000-0002-6476-4953>

Resumo

A inteligência artificial generativa é capaz de criar conteúdo a partir de bases de dados de grande volume previamente disponíveis, e o seu uso em trabalhos acadêmicos pode levar a questionamentos éticos e legais, visto que essa produção é protegida por regras, normas e leis sobre direitos autorais. Objetivo: apresentar uma análise crítica sobre o uso da inteligência artificial generativa e os direitos autorais. Método: estudo exploratório com uso de pesquisa documental numa abordagem qualitativa. Resultados: aborda conceitos que envolvem autoria e criação do espírito na produção de conteúdo por Inteligência Artificial; discute as contribuições advindas das regulamentações europeias sobre Inteligência Artificial e direitos autorais. Conclusões: observou-se a necessidade de maturação dos modelos de apropriação dos produtos gerados por Inteligência Artificial para alcance de um modelo que considere o sistema de direitos autorais para proteção das criações por Inteligência Artificial.

Palavras-chave: direitos autorais; Inteligência Artificial Generativa; modelos de apropriação.

Abstract

Generative artificial intelligence (AI) is capable of creating content based on large-scale preexisting datasets, and its use in academic work raises ethical and legal concerns, particularly because such output is subject to regulations, standards, and laws related to copyright. Objective: to present a critical analysis on the use of generative artificial intelligence and copyright. Method: exploratory study using documentary research

¹ Doutora em Ciência da Informação, Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), Brasília, DF, Brasil. Docente do Departamento de Ciência da Informação da Universidade Federal da Paraíba (UFPB), João Pessoa, Paraíba, Brasil. rosilenesousa@ibict.br.

² Doutora em Ciência da Informação, Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), Brasília, DF, Brasil. Docente da Faculdade de Medicina de Marília (FAMEMA), Marília, São Paulo, Brasil. natinakano@gmail.com.

³ Doutor em Ciência da Informação, Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), Brasília, DF, Brasil. Pesquisador, Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), Brasília, DF, Brasil. shintaku@ibict.br.

in a qualitative approach. Results: the paper addresses concepts involving authorship and creation of the spirit in content production by Artificial Intelligence; discusses the contributions arising from European regulations on Artificial Intelligence and copyright. Conclusions: the study observed the need to mature the models of appropriation of products generated by Artificial Intelligence to achieve a model that considers the copyright system to protect creations by Artificial Intelligence.

Keywords: copyright; Generative Artificial Intelligence; appropriation models.

1 INTRODUÇÃO

Historicamente, sabe-se que a Inteligência Artificial (IA) não é uma novidade, Kassan (2005), em um breve relato sobre o assunto, defende que desde o nascimento da computação a IA estava presente, mas que o marco inicial foi em 1955, com a proposta de um projeto de pesquisa. Segundo McCarthy, Minsky, Rochester e Shannon (1955), autores do projeto, era um estudo inicial de dois meses, com dez pesquisadores, com base em conjecturas, de forma a pensar como fazer que computadores possam fazer uso da linguagem, fazerem abstrações e conceitos para resolver problemas restritos aos seres humanos.

Desde então, cada vez mais estudos têm sido realizados na área de IA, com o desenvolvimento de ferramentas que implementam os resultados. Kassan (2005) descreve que a IA tem se desenvolvido em três grandes áreas, nem sempre com fronteiras bem demarcadas, nomeadamente: connexionismo, computacionalismo e robótica. Assim, dentro dos estudos do computacionalismo estão, entre outros pontos, os relacionados com a linguagem natural, que trata a mente como um sistema de símbolos formais, considerados como a linguagem do pensamento.

Neste caminho inicialmente relacionado à linguagem natural, surge a Inteligência Artificial Generativa (IAGen) relacionada com a aprendizagem de máquina, com mecanismos neocibernéticos e de feedback, em sistemas abertos que combinam formas diversificadas de técnicas de programação, em que a atuação humana pode orientar o desenvolvimento da máquina (Van der Zant; Schomaker, 2013). Por sua vez, Banh e Strobel (2023) consideram que a IAGen é um novo paradigma, em que há uma nova forma de atuar com a inteligência artificial, mudando de tarefas discriminativas e orientadas por dados para tarefas sofisticadas e criativas, capazes de gerar artefatos informacionais.

A evolução da IAGen chega a ponto que Ramos (2023) relata o uso dessa tecnologia em pesquisas acadêmicas, levantando questões a serem discutidas, ante as suas possibilidades. Nesse caso, levanta outro problema, na medida em que trabalhos acadêmicos, assim como os artísticos, estão amparados por leis e normas relacionadas aos direitos autorais. Assim, perguntas podem ser propostas como: Se parte do texto foi gerado por inteligência artificial generativa, quem é o autor? O texto gerado por IAGen deve ser considerado como uma citação? Assim, ainda há dúvidas sobre questões legais e éticas quanto a IAGen.

Nesse contexto, o presente trabalho visa a refletir sobre o uso da inteligência artificial generativa e os direitos autorais, na medida em que cada vez mais essa técnica tem sido utilizada para gerar conteúdo, inclusive acadêmicos e artísticos. Com isso, contribuir com a discussão do uso da inteligência artificial nas diversas atividades humanas cobertas pelas normas e leis que regem os direitos autorais, ou mesmo questões de cunho ético ou comercial.

1.1 Atribuição de Autoria na Lei de Direitos Autorais e na Produção de Conteúdo por Inteligência Artificial

O direito autoral tem vivido o problema causado pela sua complexidade e pela necessidade de adequações frente a interesses na proteção de lacunas que impedem a abrangência da diversidade das criações humanas.

Desse modo, surge a necessidade de proteção de uma nova categoria de bens, considerados bens informacionais autônomos criados a partir da IA. Esses bens constituem uma categoria que necessita ser regulamentada. Para tanto, busca-se evitar a desqualificação da existência dos direitos patrimoniais e morais existentes, mas direcionar o enfoque que envolve a IA, bem como a perspectiva jurídica do tema no contexto do direito privado.

O artigo 11 da Lei de Direitos Autorais (LDA) define o conceito de autor como "pessoa física criadora de obra literária, artística e científica" (Brasil, 1998). Nesse sentido, o autor, ao produzir uma criação intelectual, estabelece um vínculo com sua obra, expressando elementos que refletem sua individualidade e podem ser identificados com base em aspectos ligados à sua personalidade. A criação autoral refere-se à maneira como

a obra se manifesta, ou seja, à sua forma de expressão, que se distingue de seu suporte material.

De acordo com o artigo 11 da LDA, o autor será sempre uma pessoa física, criadora de obras literárias, artísticas ou científicas. A relação entre o autor e sua criação o caracteriza como titular originário dos direitos sobre a obra. O artigo 5º, XIV da LDA, estabelece que "considera-se titular originário o autor de obra intelectual, o intérprete, o executante, o produtor fonográfico e as empresas de radiodifusão" (Brasil, 1998).

Excepcionalmente, conforme o parágrafo único do artigo 11 da LDA, "a proteção concedida ao autor poderá aplicar-se às pessoas jurídicas nos casos previstos em lei" (Brasil, 1998). Assim, o autor, pessoa física e titular originário, pode transferir seus direitos patrimoniais sobre suas criações a terceiros, seja por meio de contrato ou sucessão, conferindo-lhes a titularidade derivada.

O artigo 14 da LDA também estabelece que o titular dos direitos de autor é aquele que adapta, traduz, arranja ou orchestra obra de domínio público. Panzolini e Demartini (2017, p. 30) destacam que "a titularidade derivada nunca abrange a totalidade dos direitos autorais (morais e patrimoniais)", reforçando que "os direitos patrimoniais são transferíveis, mas os direitos morais são inalienáveis, imprescritíveis e intransferíveis".

Torna-se fundamental lembrar que, em relação aos direitos morais, o autor sempre terá direito de paternidade sobre a obra. Esses direitos, sendo uma extensão de sua personalidade, não podem ser transferidos. A titularidade de direitos patrimoniais, por sua vez, refere-se ao direito econômico sobre a obra, que pode ser exercido pelos titulares derivados.

Nesse contexto, as criações produzidas por Inteligência Artificial (IA), apresentam desafios ao conceito legal de autoria uma vez que sua origem advém dos algoritmos e máquinas desprovidas de personalidade jurídica, e, consequentemente, no sentido legal não pode ser considerada autora em virtude da lei. Diante disso, surgem diversos questionamentos, dentre os quais, evidencia-se a quem se deve a detenção dos direitos patrimoniais de obras criadas por IA. Além disso, a quem seria atribuída a autoria em colaboração entre humanos e IA, ou ainda, texto gerado por IAGen deve ser considerado como uma criação; e, qual o reflexo das criações de IA na ética.

1.2 Inteligência Artificial e Direitos Autorais

Conceitualmente, segundo Russell e Norvig (2021, p. 2), pode-se observar oito definições de Inteligência Artificial, que geralmente se relacionam a processos de pensamento e raciocínio, comportamento, desempenho humano e racionalidade. Nessa perspectiva, os conceitos apresentam-se conforme exposto na Figura 1.

Figura 1 - Algumas definições de Inteligência Artificial, organizadas em quatro categorias

Pensando como um humano	Pensando racionalmente
"O novo e interessante esforço para fazer os computadores pensarem (...) <i>máquinas com mentes</i> , no sentido total e literal." (Haugeland, 1985)	"O estudo das faculdades mentais pelo uso de modelos computacionais." (Charniak e McDermott, 1985)
"[Automatização de] atividades que associamos ao pensamento humano, atividades como a tomada de decisões, a resolução de problemas, o aprendizado..." (Bellman, 1978)	"O estudo das computações que tornam possível perceber, raciocinar e agir." (Winston, 1992)
Agindo como seres humanos	Agindo racionalmente
"A arte de criar máquinas que executam funções que exigem inteligência quando executadas por pessoas." (Kurzweil, 1990)	"Inteligência Computacional é o estudo do projeto de agentes inteligentes." (Poole <i>et al.</i> , 1998)
"O estudo de como os computadores podem fazer tarefas que hoje são melhor desempenhadas pelas pessoas." (Rich and Knight, 1991)	"AI... está relacionada a um desempenho inteligente de artefatos." (Nilsson, 1998)

Fonte: Russell e Norvig (2021, p. 2).

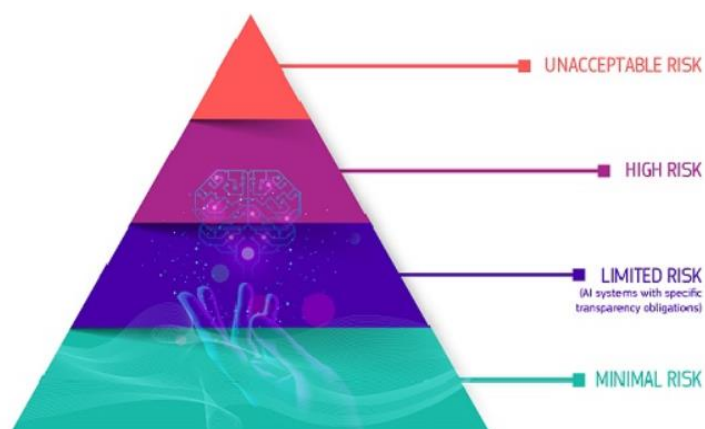
A Norma ISO ABNT NBR ISO/IEC 22989:2023, que trata de conceitos de Inteligência Artificial e terminologia conceitua os sistemas de inteligência artificial como aquele desenvolvido para gerar saídas como conteúdo, previsões, recomendações ou decisões para um determinado conjunto de objetivos definidos pelo homem que pode utilizar diversas técnicas de abordagem para desenvolvimento de modelo, representação de dados, conhecimentos processos, dentre outros, para realização de tarefas operando com diferentes níveis de automação (Associação Brasileira de Normas Técnicas, 2023, p. 2). A referida norma ainda evidencia que não se deve confundir a IA generalista da IA generativa. A IA generalista, ainda impossível atualmente, refere-se a sistemas de IA projetados para realizar uma ampla variedade de tarefas, imitando a inteligência humana, capaz de aprender, raciocinar, resolver problemas e se adaptar a diferentes contextos. Por outro lado, a IA generativa, popular nos dias de hoje e foco da análise deste estudo, está direcionada à criação de novos conteúdos, como imagens, texto, música, ou outros tipos de dados, com base em padrões aprendidos de grandes conjuntos de dados. Assim, enquanto a IA generalista busca replicar a capacidade humana de realizar múltiplas tarefas, a IA generativa se especializa na produção de novos conteúdos dentro de contextos específicos.

Nesse sentido, surge a necessidade de regulação de seu uso em face de seu vertiginoso crescimento. A União Europeia em abril de 2021 propôs inicialmente um

texto, seguido pelo do Conselho em 2022, e do texto do Parlamento Europeu em junho de 2023, resultando, a partir do trílogo entre Parlamento Europeu, Conselho da União Europeia e a Comissão Europeia, no Regulamento (UE) 2024/1689 denominado também de AI Act.

O *AI Act*, que estabelece regras harmonizadas sobre inteligência artificial, surge como primeiro arcabouço legal sobre inteligência artificial, com abordagem sobre riscos potencializando a Europa no papel de liderança global (Comissão Europeia, 2024). O *AI Act* centra-se nos sistemas considerados de alto risco, definindo assim quatro níveis de risco, conforme apresentado na Figura 2.

Figura 2 - Quatro níveis de risco para sistemas de IA



Fonte: Comissão Europeia (2024).

A proposta brasileira toma como base os níveis de risco conforme adotado pela UE. Segundo exposto por Schirru e Souza (2024), o *AI Act* reflete o amadurecimento das discussões entre a IA generativa e os direitos autorais.

Com relação à Lei de Direitos Autorais (LDA) vigente no Brasil, observa-se que o status jurídico das criações por meio de IA sem interferência humana relevante se configura como uma lacuna importante do cenário jurídico brasileiro, e, portanto, alguns modelos estão sendo discutidos pelos operadores do direito.

3 PROCEDIMENTOS METODOLÓGICOS

Trata-se de estudo exploratório com uso de pesquisa documental numa abordagem qualitativa. Conduz uma análise crítica com base na pesquisa bibliográfica e

na exploração de documentos legais relevantes, como a Lei de Direitos Autorais (Brasil, 1998). Quanto à pesquisa bibliográfica, concentra-se principalmente nos estudos teóricos de Schirru (2020) e Wachowicz (2021). A seleção desses autores se deu em face do pioneirismo da discussão que conduzem sobre o tema. Examina-se como o regime jurídico brasileiro trata a questão dos direitos de propriedade intelectual relacionados a modelos de apropriação de produtos gerados por sistemas de IA para compreender e avaliar as diversas perspectivas e propostas por Schirru (2020) e Wachowicz (2021). Esses autores oferecem percepções importantes para refletir sobre a legislação vigente e as implicações legais dos direitos autorais em contextos influenciados pela IA, servindo como a espinha dorsal para o desenvolvimento de argumentos críticos, propostos neste estudo.

4 RESULTADOS

Esta seção apresenta os resultados da pesquisa documental e bibliográfica do estudo, apresentando os modelos de apropriação de produtos gerados por inteligência artificial discutidos a partir dos trabalhos dos autores pioneiros selecionados.

Os modelos de apropriação de produtos gerados por IA buscam responder à questão de como os produtos gerados por IA devem ser tratados no sistema jurídico, especialmente em relação à titularidade e proteção autoral. Pautados nos estudos conduzidos por Schirru (2020) e por Wachowicz (2021) realiza-se uma análise crítica para discorrer sobre tais modelos.

No Brasil, o debate sobre a titularidade de obras geradas por inteligência artificial (IA) está dividido em quatro principais correntes (Wachowicz, 2021). Cada corrente oferece uma solução distinta para lidar com as complexidades da criação autônoma de IA e sua relação com o direito autoral.

A primeira corrente, pautada no art. 45 da LDA (Brasil, 1998), sustenta que as obras geradas por IA deveriam ser automaticamente colocadas em domínio público:

Art. 45. Além das obras em relação às quais decorreu o prazo de proteção aos direitos patrimoniais, pertencem ao domínio público:
I - as de autores falecidos que não tenham deixado sucessores;
II - as de autor desconhecido, ressalvada a proteção legal aos conhecimentos étnicos e tradicionais.

O argumento central consiste no entendimento de que, como os direitos autorais implicam em uma criação humana, e os algoritmos não possuem capacidade de intenção ou criatividade, não seria apropriado recompensar uma criação que não envolve o trabalho criativo humano. Assim, essas obras estariam disponíveis para uso livre, pois não podem ser associadas a um autor humano que detém direitos sobre elas.

A segunda corrente defende que a titularidade das criações de IA deve ser atribuída à empresa que desenvolveu o software ou tecnologia. Isso se justifica por aspectos de propriedade intelectual, investimentos feitos no desenvolvimento da IA, e o controle que a empresa possui sobre a tecnologia. Além disso, atribuir a titularidade à empresa permitiria que ela controlasse como as obras são usadas e geradas, além de ter a responsabilidade legal sobre possíveis questões envolvendo essas criações (Wachowicz, 2021).

A terceira corrente advoga que a atribuição de titularidade pertence ao usuário da IA, uma vez que ele utiliza o sistema para gerar as obras. Essa corrente defende que, embora a IA seja responsável pelo processo criativo, o usuário é quem define os parâmetros, toma as decisões iniciais e fornece as instruções que guiam a IA na criação. Assim, o usuário tem papel ativo e criativo no processo, o que justificaria a atribuição dos direitos autorais (Wachowicz, 2021).

A última corrente defende a criação de um novo direito conexo aos direitos autorais. Esse modelo visa recompensar economicamente as empresas que desenvolvem tecnologias de IA, sem necessariamente aplicar o regime tradicional de direitos autorais personalíssimo, porém, evitando conflitar o princípio fundamental dos direitos autorais de que a criatividade deve ser humana (Wachowicz, 2021).

Wachowicz (2021) enfatiza que, embora as inovações tecnológicas ofereçam grandes oportunidades, elas também trazem desafios para o sistema jurídico, especialmente no que diz respeito à autoria e titularidade das obras geradas por IA. Ele destaca a necessidade urgente de regulamentações claras que possam enfrentar essas complexidades e preencher o atual vácuo legislativo. As criações por IA colocam em questão conceitos tradicionais de direitos autorais, exigindo novas abordagens para garantir que tanto a proteção da criatividade humana quanto o incentivo à inovação tecnológica sejam adequadamente equilibrados.

Em linha com os estudos desenvolvidos por Wachowicz (2021), Schirru (2020), em seus estudos sobre direitos autorais e inteligência artificial (IA), discute diferentes modelos de apropriação das criações geradas por IA e, portanto, como o ordenamento jurídico brasileiro poderia tratar a titularidade e proteção autoral desses produtos.

Os *Modelos Antropocêntricos* são centrados no ser humano. A titularidade, portanto, é sempre atribuída a uma pessoa física, como o programador ou o operador que controla a IA, descartando qualquer reconhecimento da IA como autora. O direito autoral é entendido como um direito personalíssimo, inerente ao ser humano.

A segunda categoria de modelos proposta por Shirru (2020) intitula-se *Modelos Centrados no Titular*. Shirru (2020) avança para os modelos que tratam os produtos de IA equiparados às obras feitas sob encomenda; à proteção por direitos conexos, e o tratamento equiparado a obras coletivas.

Com relação aos Modelos Tecnocêntricos (Shirru, 2020), esses exploram a ideia de atribuir autoria ou titularidade de direitos autorais diretamente a sistemas de inteligência artificial (IA), sem associá-los a um ser humano ou a uma pessoa jurídica. Esses modelos propõem uma ruptura com o paradigma tradicional antropocêntrico em direito autoral, que sempre centralizou os direitos de criação e propriedade intelectual em indivíduos ou entidades humanas. Na proposta, os sistemas de IA são considerados titulares. Shirru (2020, p. 300) destaca:

- (i) a criação de um sistema de classificação para robôs;
- (ii) o desenvolvimento de um sistema de registro de robôs, com a possibilidade de coordenação por uma Agência específica;
- (iii) a criação de um sistema de seguros para garantir às vítimas a indenização cabível,
- (iv) a possibilidade da criação de uma personalidade específica para esses robôs.

Propõe-se um sistema de classificação para robôs e um registro mantido por uma agência específica para padronizar o entendimento sobre diferentes tipos de robôs e sistemas de IA com base em sua autonomia e capacidades técnicas. Contudo, observa-se uma sugestão inovadora com a criação de uma personalidade eletrônica para robôs, equiparada a uma personalidade jurídica. Essa solução abordaria não apenas a responsabilidade civil, mas também a possibilidade de gestão de direitos autorais.

A integração dessa proposta nos sistemas de direitos autorais existentes revela desafios importantes que demandam uma reflexão profunda e ajustes no próprio conceito de autoria e titularidade.

Shirru (2020) analisa a aplicabilidade do conceito de *domínio público* aos produtos criados pela IA, destacando as limitações do enquadramento desses produtos dentro das normas tradicionais de direitos autorais. Argumenta-se que, devido à ausência de originalidade subjetiva, característica central para a atribuição de direitos autorais conforme definido pela legislação vigente, os produtos de IA não satisfazem os critérios necessários para serem considerados obras protegidas. Isso por causa da natureza da criação de IA, que não tem a expressão individual que define a autoria humana. Assim, propõe-se que tais produtos sejam classificados como domínio público, permitindo o uso irrestrito pela sociedade sem as restrições impostas pelos direitos autorais.

Uma última categoria de modelos para apropriação de produtos criados por IA são intitulados por Shirru (2020) como *Sui Generis* (Externos ao Sistema de Direitos Autorais). O referido autor aborda a possibilidade de desenvolver um regime *sui generis*, ou seja, um sistema legal específico e independente, alternativo ao sistema tradicional de direitos autorais. Este modelo não estaria vinculado aos princípios fundamentais do direito autoral, que requerem a presença de originalidade e autoria humana para a concessão de proteção.

Um regime *sui generis* permitiria a criação de normas ajustadas especificamente às características e necessidades dos produtos de IA, sem forçar a adaptação desses produtos a um sistema que não foi desenhado para considerá-los.

Finalmente, o autor propõe uma reflexão sobre como os produtos de IA podem ser tratados de maneira inovadora e específica através de um regime legal *sui generis*, que respeita a especificidade dessas criações e oferece uma alternativa viável ao enquadramento inadequado oferecido pelo sistema de direitos autorais tradicional.

5 CONSIDERAÇÕES FINAIS

Da análise pontual sobre o tema foi possível destacar que a evolução da IA generativa e sua capacidade de criar conteúdo autônomo suscitam questões relevantes

no âmbito dos direitos autorais. Embora a inteligência artificial tenha o potencial de enriquecer a produção acadêmica e criativa, evidencia-se uma lacuna significativa no que diz respeito à proteção jurídica dessas criações.

A Diretiva de Direitos Autorais da UE (2019/790) e o AI Act. representam normativas que visam regular o uso da IA no contexto de direitos autorais, porém, essas regulamentações ainda não são suficientes para abordar todos os desafios emergentes. Assim, destaca-se a necessidade de amadurecer os modelos de apropriação dos produtos gerados por IA. Alguns estudiosos da área sugerem um modelo *sui generis*, que envolve tanto os direitos de autor como os conexos, considerando-o solução mais adequada para equilibrar a proteção das obras e o incentivo à inovação.

Diante dos modelos de apropriação tratados, conclui-se que o desenvolvimento de um arcabouço jurídico específico para as criações de IA, que considere tanto os direitos de autor quanto os direitos conexos (*sui generis*), torna-se essencial para garantir a segurança jurídica, promover a equidade nas relações entre criadores humanos e produtor autônomos gerados por máquinas, e fomentar um ambiente de inovação sustentável no qual as contribuições tecnológicas possam coexistir com a proteção da propriedade intelectual.

Essa lacuna legislativa aponta para a necessidade de ajustes no ordenamento jurídico brasileiro, especialmente à medida que a tecnologia avança e a IA se torna cada vez mais capaz de gerar obras que, à primeira vista, parecem ser o produto de um processo criativo, mas que, em essência, carecem do componente humano essencial à proteção autoral. Dessa forma, a recomendação para legisladores, criadores de conteúdo, desenvolvedores de IA, acadêmicos, instituições governamentais, entre outros, seria considerar os aspectos éticos que cada campo ou área atualmente determina como práticas éticas ou boas práticas no uso da IA.

Para estudos futuros, recomenda-se um estudo direcionado às IA e sobre como os seus criadores e desenvolvedores alimentam suas bases de dados, considerando os direitos autorais.

REFERÊNCIAS

- ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **ABNT NBR ISO/IEC 22989:2023**: tecnologia da informação — inteligência artificial — conceitos de inteligência artificial e terminologia. Rio de Janeiro: ABNT, 2023.
- BANH, Leonardo; STROBEL, Gero. Generative artificial intelligence. **Electronic Markets**, v. 33, n. 1, p. 63, 2023.
- BRASIL. Lei nº 9.610, de 19 de fevereiro de 1998. Altera, atualiza e consolida a legislação sobre direitos autorais e dá outras providências. **Diário Oficial da União**. Brasília, DF: Casa Civil, 1998. Disponível em: http://www.planalto.gov.br/ccivil_03/leis/L9610.htm. Acesso em: 25 out. 2023.
- COMISSÃO EUROPEIA. **IA Act**. 2024. Disponível em: <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>. Acesso em: 14 out. 2024.
- KASSAN, Peter. AI gone awry: the futile quest for artificial intelligence. **Skeptic**: Altadena, v. 12, n. 2, p. 30-40, 2005. Disponível em: https://www.skeptic.com/reading_room/artificial-intelligence-gone-awry/. Acesso em: 14 out. 2024.
- MCCARTHY, John *et al.* A proposal for the dartmouth summer research project on artificial intelligence, 1955. **AI magazine**, v. 27, n. 4, p. 12-12, 2006. Disponível em: <https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1904>. Acesso em: 14 out. 2024.
- PANZOLINI, Carolina; DEMARTINI, Silvana. **Direitos autorais**: perguntas e respostas. Brasília: Secretaria-Geral de Administração, 2017. Disponível em: https://portal.tcu.gov.br/data/files/10/01/F5/3A/AE91F6107AD96FE6F18818A8/Manual_direitos_autorais_FAQ.pdf. Acesso em: 12 jan. 2021.
- RAMOS, A. S. M. Generative artificial intelligence based on large language models: tools for use in academic research. **Scielo Preprints**. 2023. Disponível em: <https://doi.org/10.1590/SciELOPreprints.6105>. Acesso em: 14 out. 2024.
- RUSSELL, Stuart J.; NORVIG, Peter. **Inteligência artificial**. 3. ed. Rio de Janeiro: LTC, 2021.
- SCHIRRU, Luca; SOUZA, Allan Rocha de. IA, direitos autorais e o AI act: o que nos diz sobre regulação. **Jota**, 2024. Disponível em: <https://www.jota.info/artigos/ia-direitos-autorais-e-o-ai-act-o-que-nos-diz-sobre-regulacao>. Acesso em: 14 out. 2024.
- SCHIRRU, Luca. **Direito autoral e inteligência artificial**: autoria e titularidade nos produtos de IA. Tese (doutorado) – Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2020. Disponível em: https://www.ie.ufrj.br/images/IE/PPED/Teses/2020/Vers%C3%A3o%20final_Tese%20Luca_PDFa.pdf. Acesso em: 14 out. 2024.

UNIÃO EUROPEIA. Parlamento Europeu. Conselho. Regulamento (UE) 2024/1689 do Parlamento Europeu e do Conselho, de 13 de março de 2024. Regulamentação de inteligência artificial. **Jornal Oficial da União Europeia**, Luxemburgo, L 1689, 13 mar. 2024. Disponível em: https://eur-lex.europa.eu/legal-content/PT/TXT/PDF/?uri=OJ:L_202401689. Acesso em: 14 out. 2024.

UNIÃO EUROPEIA. Resolução do Parlamento Europeu, de 16 de fevereiro de 2017. **Recomendações à comissão sobre disposições de direito civil sobre robótica**. 2015. Disponível em: http://www.europarl.europa.eu/doceo/document/TA-8-2017-0051_PT.pdf. Acesso em 14 out. 2024.

VAN DER ZANT, Tijn; KOUW, Matthijs; SCHOMAKER, Lambert. **Generative artificial intelligence**. Springer Berlin: Heidelberg, 2013.

WACHOWICZ, Marcos, *et al.* **Direito autoral & inteligência artificial**. Curitiba: Ioda, 2021.



[4.0 Internacional](https://creativecommons.org/licenses/by/4.0/)

Esta licença permite compartilhamento, remixe, adaptação e criação a partir do trabalho, mesmo para fins comerciais, desde que sejam atribuídos créditos ao(s) autor(es). Conteúdos elaborados por terceiros, citados e referenciados nesta obra não são cobertos pela licença.

Como citar este trabalho:

SOUSA, Rosilene Paiva Marinho de; NAKANO, Natalia; SHINTAKU, Milton. Reflexões sobre inteligência artificial generativa e os direitos autorais. *In*: WORKSHOP DE INFORMAÇÃO, DADOS E TECNOLOGIA, 8., Marília, SP. **Anais [...]**. Marília, SP: Universidade de Marília. 2025. DOI: <https://doi.org/10.22477/viii.widat.243>.



WIDaT 2025

VIII WORKSHOP DE INFORMAÇÃO,
DADOS E TECNOLOGIA

MARÍLIA • SP

Aprendizado de Máquina para detecção de sites de *Phishing*

Machine Learning for Phishing website detection

Pedro Henrique Del Bianco Hokama¹

 <http://lattes.cnpq.br/9395192779202995>

 <https://orcid.org/0000-0002-3734-7679>

Jader Duque Figueredo²

 <http://lattes.cnpq.br/5323391336547899>

 <https://orcid.org/0009-0005-6281-3386>

Karen Cristina Soares Cavalheiro³

 <http://lattes.cnpq.br/3606856318278865>

 <https://orcid.org/0000-0003-2513-6927>

Resumo

No meio digital, o termo Phishing se refere a materiais que simulam algum serviço com o objetivo de enganar os usuários para que enviem informações pessoais. Este artigo apresenta a utilização da técnica de árvores interpretáveis para a detecção de sites de Phishing. Essas árvores têm a característica de serem compreensíveis e possivelmente utilizadas por usuários comuns, mesmo sem o auxílio de um software especializado. Foi utilizada uma formulação em Programação Linear Inteira para encontrar árvores que classificam um conjunto de treinamento da melhor forma possível. Resultados parciais apontam que observando poucos elementos sobre um site foi possível detectar um Phishing com 87% de precisão no conjunto de testes.

Palavras-chave: golpes digitais; documentos falsos; árvores de classificação.

Abstract

In the digital realm, the term Phishing refers to materials that simulate a legitimate service with the aim of deceiving users into submitting personal information. We present the approach of interpretable tree techniques for phishing website detection. These trees possess the characteristic of being potentially usable by everyday users, even without the aid of specialized software. We employ an Integer Linear Programming formulation to find trees that classify a training set in the best possible way. Partial results indicate that by observing a few elements about a website, it was possible to detect phishing with 87% accuracy in the test set.

Keywords: digital scams; fraudulent documents; classification trees.

¹ Doutor em Ciência da Computação, Universidade Estadual de Campinas (UNICAMP), Campinas, SP, Brasil. Professor do Magistério Superior, Universidade Federal de Itajubá (UNIFEI), Itajubá, MG, Brasil. Hokama@unifei.edu.br.

² Mestre em Ciência e Tecnologia da Computação, Universidade Federal de Itajubá (UNIFEI), Itajubá, MG, Brasil. Analista de Tecnologia da Informação, Universidade Federal de Itajubá (UNIFEI), Itajubá, MG, Brasil. jader.figueiredo@unifei.edu.br.

³ Mestre em Ciência da Informação, Universidade Federal de São Carlos (UFSCar), São Carlos, SP, Brasil. Doutoranda, Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP), Marília, SP, Brasil. karen.cavalheiro@unesp.br.

1 INTRODUÇÃO

A crescente dependência da sociedade moderna em tecnologias digitais tem sido acompanhada por um aumento significativo nas ameaças cibernéticas, com destaque para a prática danosa de *phishing* (Khonji; Iraqi; Jones, 2013). O *phishing* é uma forma de ataque em que indivíduos mal-intencionados tentam enganar usuários, induzindo-os a revelar informações confidenciais, como senhas e dados pessoais, por meio de mensagens falsas e sites fraudulentos. Para combater essa ameaça, é importante utilizar técnicas eficazes de detecção de *phishing*, mas também conscientizar os usuários da internet de como evitar esse tipo de golpe.

Nesse contexto, as árvores de classificação interpretáveis emergem como uma abordagem promissora para identificar sites suspeitos associados a atividades de *phishing*. Técnicas mais comuns de aprendizado de máquinas, como as redes neurais e redes neurais profundas, têm sido largamente utilizadas com sucesso em diversas aplicações (Sharifani; Amini, 2023). Entretanto, essas técnicas sofrem da falta de interpretabilidade dos resultados, ou seja, um usuário comum ao olhar uma rede neural não consegue, por exemplo, entender o que faz um dado ser classificado de uma forma ou de outra, tão pouco conseguiria classificar uma entrada sem um programa de computador.

Em contrapartida as árvores de classificação (Bertsimas; Orfanoudaki; Wiberg, 2021) oferecem a vantagem de serem facilmente interpretáveis, permitindo que, tanto usuários comuns quanto especialistas em segurança, compreendam, validem e apliquem as decisões tomadas pelo modelo. Essa interpretabilidade é fundamental para ganhar a confiança dos usuários e criar um sistema de detecção simples. Estudos recentes destacaram a efetividade do modelo exato aplicado a decisão (Figueredo, 2024).

A técnica de Árvores de Classificação Ótimas, tem ganhado novos estudos como as Árvores de Classificação Ótimas Fortes (*Strong Optimal Classification Trees*) (Aghaei, Gómez e Vayanos, 2024) que gera árvores mais justas através da relaxação linear mais forte e uso de método de decomposição de Benders, obtendo aplicações mais práticas e poderosas de árvores de classificação.

Já a Árvore de Decisão Ótima usando Programação Dinâmica de Van Der Linden, De Weerd e Demirović (2023) não só é capaz de lidar com grande quantidade de restrições e objetivos como ainda superar as linhas de base de eficiência e escalabilidade de abordagens baseadas em Otimização Inteira Mista.

Neste trabalho é explorado o uso de árvores de classificação como uma ferramenta para detectar *sites* de *phishing*. Serão resumidos os fundamentos teóricos por trás das árvores de classificação, explicando como elas funcionam e como podem ser aplicadas à detecção de *phishing*. Também será apresentado um modelo matemático em Programação Linear Inteira (PLI) capaz de encontrar as melhores árvores possíveis para um dado conjunto de treinamento.

Safi e Singh (2023) realizaram uma revisão sistemática da literatura sobre as diferentes abordagens de detecção de *phishing*, para isso foram analisados 80 trabalhos, dentre eles artigos científicos e trabalhos apresentados em eventos ou *sites* renomados dos últimos cinco anos; os autores apontam que as técnicas de aprendizado de máquina foram as mais aplicadas, sendo utilizadas em 57 estudos e o algoritmo que atingiu maior precisão na detecção de *sites* de *phishing* foi a Rede Neural Convolucional, com 99,98% de acerto.

O objetivo geral desta pesquisa é auxiliar usuários a detectarem se o *site* em que estão navegando possui *phishing*. O objetivo específico é: elencar orientações simples e capazes de ser identificadas com boa precisão que possam classificar um site como *phishing* ou não.

A escolha do modelo em Programação Linear Inteira se deve ao fato de que, para realizar a classificação ele tem mostrado bons resultados em trabalhos recentes (Bertsimas; Orfanoudaki; Wiberg, 2021; Figueredo, 2024). Destaca-se ainda que esse é um trabalho parcial, em trabalhos futuros os modelos podem ser expandidos e testados em diferentes aplicações.

Este trabalho está dividido em seções, são elas: a seção 2, que é apresentada a metodologia empregada; a seção 3, que contempla o desenvolvimento e a fundamentação teórica da árvore de classificação e da PLI; A seção 4, que apresenta os

resultados parciais obtidos até o momento e por fim, a seção 5, na qual constam as conclusões dos autores e trabalhos futuros.

2 PROCEDIMENTOS METODOLÓGICOS

Para o desenvolvimento da fundamentação teórica da pesquisa foi realizado um levantamento documental sobre os temas propostos: *Phishing* golpes digitais; documentos falsos; árvores de classificação.

A pesquisa utiliza o método Árvore de Classificação Ótima e se caracteriza como quali-quantitativa, de caráter descritivo e exploratório, sobre os temas de: Aprendizado de máquina, detecção de informação falsa e golpes e modelagem matemática.

Para realização da pesquisa os modelos foram implementados utilizando a linguagem de programação C++, compilado em g++ 11.3.0 e executado em ambiente Linux. O resolvidor de Programação Linear Inteira utilizado foi o CPLEX 20.1.

Os experimentos foram realizados com o *dataset "Phishing website Detector"* de Eswar Chand (2020) disponível na plataforma *Kaggle*. O *dataset* é formado por 11054 observações com 30 atributos (variáveis dependentes) e 1 classe (variável dependente), a qual pretende-se prever, organizados em tabela, onde as observações são as linhas e os atributos são as colunas. Os atributos são todos numéricos, onde o valor "1" significa que o atributo ocorre, "0" significa neutro ou incerto e "-1" significa que o atributo não acontece. Todas as observações são classificadas como "1" seguro, ou "-1" malicioso. O conjunto foi dividido em 2, onde a primeira metade foi usada para treinamento e a outra metade para teste.

3 DESENVOLVIMENTO

Árvores de classificação interpretáveis são modelos de aprendizado de máquina que representam um conjunto de regras de decisão hierárquicas em uma estrutura de árvore. Cada nó interno da árvore representa uma condição sobre um atributo, enquanto os nós folha representam as classes ou rótulos de saída. Essas árvores são ditas interpretáveis, pois permitem uma compreensão intuitiva das regras de classificação

utilizadas pelo modelo. Os caminhos na árvore representam as sequências de decisões tomadas para chegar a uma classificação específica, facilitando a interpretação e a explicação do processo de tomada de decisão.

Os usuários podem utilizar as árvores da seguinte forma: no topo da árvore existe uma pergunta, caso a pergunta seja respondida com “não”, o usuário segue a linha/galho da direita, caso a resposta seja “sim”, segue para a esquerda, e continua para as seguintes perguntas até que no final seja respondido se o Localizador Uniforme de Recursos (URL) é legítimo ou não, e o usuário saberá se está sujeito à *phishing*; se a resposta final respondida for “legítimo”, significa que é seguro de ser acessado.

A interpretabilidade das árvores de classificação é um atributo valioso em diversas aplicações, especialmente em problemas nos quais é desejável que a compreensão do modelo possa ser utilizada como instruções simples, que um humano possa seguir sem o auxílio de um *software*. A capacidade de examinar as regras de decisão em cada nó permite que especialistas em segurança compreendam quais características ou indicadores são relevantes para identificar um *site* malicioso. Além disso, as árvores de classificação interpretáveis permitem que os usuários finais entendam por que um *site* foi classificado como suspeito ou seguro, aumentando a confiança e a transparência do sistema de detecção.

3.1 Treinamento e Teste

O conjunto de treinamento é uma parte dos dados disponíveis usados para ajustar um modelo de aprendizado de máquina. Ele consiste em exemplos de entradas (por exemplo: *sites*) associados às respectivas saídas desejadas (por exemplo: se é um *site* de *phishing* ou não). O objetivo do treinamento é permitir que o modelo aprenda padrões e relações entre os dados de entrada e saída, ajustando seus parâmetros e melhorando seu desempenho ao longo do tempo. O conjunto de treinamento é fundamental para criar um modelo preciso e geral, capaz de fazer previsões precisas em novos dados.

Por outro lado, o conjunto de teste é uma porção separada dos dados que não é usada durante o treinamento, mas sim para avaliar o desempenho do modelo após o treinamento. Esse conjunto de teste é composto por exemplos que não foram

apresentados ao modelo durante o treinamento, permitindo verificar a capacidade de generalização do modelo para dados não vistos anteriormente. Ao avaliar o modelo com o conjunto de teste, pode-se obter métricas de desempenho, como acurácia, que indicam quão bem o modelo está se saindo na detecção de *phishing* não observadas anteriormente. Isso ajuda a estimar o desempenho do modelo em cenários reais e a verificar se ele é capaz de generalizar efetivamente para novos dados.

3.2 Fundamentos da Programação Linear Inteira

A Programação Linear Inteira (Wolsey, 2020) é uma técnica poderosa utilizada para otimizar problemas em que as variáveis de decisão devem assumir valores inteiros. Ela combina conceitos da Programação Linear, que lida com variáveis contínuas, com a restrição adicional de que as variáveis devem ser inteiras. A PLI tem sido, por décadas (Wolsey; Nemhauser, 1999), amplamente aplicada em diferentes áreas, como logística, produção e pesquisa operacional.

Ao utilizar PLI para encontrar árvores de classificação interpretáveis, o problema de construção da árvore é formulado como um problema de otimização, em que o objetivo é encontrar a melhor árvore de classificação com base em critérios específicos, como a acurácia de classificação e a interpretabilidade das regras geradas (complexidade da árvore).

Informalmente, um “Programa Linear Inteiro” é um modelo matemático que descreve como uma solução para o problema deve ser, ou seja, quais restrições ela deve respeitar relacionadas a uma função objetivo que descreve o que deve ser otimizado. Esse modelo então é resolvido por um algoritmo e obtém-se a melhor solução possível.

3.3 Formulação do Problema de Árvores de Classificação com PLI

Para utilizar PLI na construção de árvores de classificação interpretáveis para detecção de *phishing* em URLs, o problema é formulado como um modelo de otimização. O modelo é composto por variáveis de decisão que representam as regras de divisão em

cada nó da árvore e as relações entre os nós. O objetivo é maximizar a acurácia de classificação da árvore, enquanto se impõem restrições relacionadas à interpretabilidade.

As restrições de interpretabilidade podem ser definidas de várias formas, dependendo dos critérios desejados. Por exemplo, pode-se limitar o número máximo de níveis na árvore, ou ainda punindo a quantidade de divisões, fazendo com que a solução apenas as crie quando valerem a pena. Essas restrições são essenciais para garantir que a árvore resultante seja compreensível e interpretável.

O modelo apresentado a seguir foi adaptado de Bertsimas, Orfanoudaki e Wiberg (2021), primeiramente serão definidos os dados do problema, depois as variáveis, e por fim o modelo. São dados do problema:

- a) I é o conjunto dos dados de treinamento. Cada elemento i de I é um *site*;
- b) P é o conjunto de atributos, e.g., (Usa Subdomínio, usa encurtador de URL e etc);
- c) K é o conjunto de classes {*Phishing*, Legítimo};
- d) x_{ip} é o valor do parâmetro p para o item $i \in I$;
- e) t é cada nó dentre um conjunto T de nós da árvore;
- f) T_B é o conjunto de nós internos da árvore;
- g) T_L é o conjunto de nós folhas da árvore;
- h) $A_R(t)$ são os ancestrais do nó t onde foi necessário descer à esquerda para chegar em t ;
- i) $A_L(t)$ são os ancestrais do nó t onde foi necessário descer à direita para chegar em t ;

As variáveis do modelo são:

- a) a_{tp} indica se o parâmetro p é usado para a divisão no nó t ;
- b) b_t indica qual o valor que faz um elemento ir para direita ou para esquerda;
- c) z_{it} indica se o item i está na folha t ;
- d) c_{kt} indica se a folha t está classificada como classe k ;
- e) N_{kt} indica a quantidade de elementos da classe k que estão na folha t ;
- f) N_t indica a quantidade total de elementos na folha t ;
- g) L_t indica a quantidade de erros na folha t .

O modelo então é apresentado a seguir e posteriormente uma intuição do que significa cada uma das restrições.

$$\text{Min. } \sum_{t \in \mathcal{T}_L} L_t, \quad (1)$$

$$\text{s. a } L_t \geq N_t - N_{kt} - |I|(1 - c_{kt}), \quad k \in K, \forall t \in \mathcal{T}_L, \quad (2)$$

$$N_{kt} = \sum_{i \in I: y_i = k} z_{it}, \quad k \in K, \forall t \in \mathcal{T}_L, \quad (3)$$

$$N_t = \sum_{i \in I} z_{it}, \quad \forall t \in \mathcal{T}_L, \quad (4)$$

$$\sum_{k \in K} c_{kt} = 1, \quad \forall t \in \mathcal{T}_L, \quad (5)$$

$$\sum_{p \in P} a_{mp} x_{ip} \geq b_m - M(1 - z_{it}), \quad i \in I, \forall t \in \mathcal{T}_L, \forall m \in A_R(t), \quad (6)$$

$$\sum_{p \in P} a_{mp} x_{ip} + \epsilon \leq b_m + M(1 - z_{it}), \quad i \in I, \forall t \in \mathcal{T}_L, \forall m \in A_L(t), \quad (7)$$

$$\sum_{p \in P} a_{tp} = 1, \quad \forall t \in \mathcal{T}_B, \quad (8)$$

$$0 \leq b_t \leq 1, \quad \forall t \in \mathcal{T}_B, \quad (9)$$

$$z_{it} \in \{0, 1\}, \quad i \in I, \forall t \in \mathcal{T}_L, \quad (10)$$

$$c_{kt} \in \{0, 1\}, \quad k \in K, \forall t \in \mathcal{T}_L, \quad (11)$$

$$a_{tp} \in \{0, 1\}, \quad p \in P, \forall t \in \mathcal{T}_B, \quad (12)$$

A função objetivo (1) busca minimizar o número total de erros. A restrição (2) faz com que o número de erros em cada folha seja corretamente computado. A restrição (3) contabiliza o número de elementos de cada classe em cada folha. A restrição (4) computa a quantidade total de elementos por folha. A restrição (5) obriga toda folha a ser classificada como uma classe. As restrições (6) e (7) juntas fazem com que o elemento, para chegar em uma certa folha, tenha percorrido adequadamente as divisões dos ancestrais. A restrição (8) obriga todo galho a escolher um parâmetro para a divisão. Por fim as restrições (9), (10), (11) e (12) indicam o domínio de cada variável.

4 RESULTADOS PRELIMINARES

Como o número de variáveis do modelo depende do número de exemplares no conjunto de treinamento, e esse número impacta diretamente no tempo de resolução do modelo, optou-se por limitar o número de exemplares do conjunto de treinamento aos 150 primeiros, e o conjunto de testes são os 150 seguintes.

O modelo apresentado na seção 3 foi implementado em linguagem C++, compilado com g++ 11.3.0 e executado em ambiente Linux. O resolvidor de Programação Linear Inteira utilizado foi o CPLEX 20.1.

Para reduzir o número de parâmetros e preservar a interpretabilidade das árvores geradas, eliminou-se do *dataset* todos os parâmetros que um usuário leigo pudesse ter dificuldade de verificar. Dessa forma os resultados obtidos podem ser utilizados como instrução para a detecção de *phishing* mesmo sem o uso de softwares adicionais. Vale notar que se não houvesse o objetivo de tornar a árvore aplicável por um usuário, poderíamos usar um método como o *Principal Component Analysis* - PCA (Greenacre *et al.*, 2022) para reduzir o número de parâmetros. Ao final sobraram 11 parâmetros:

- a) 1. A URL do *site* usa um endereço de protocolo de internet (IP)?
- b) 2. A URL é demasiadamente longa?
- c) 3. O *site* usa encurtador de URL?
- d) 4. A URL usa o símbolo de @?
- e) 5. O domínio tem um traço/hífen “-”?
- f) 6. A URL tem subdomínios?
- g) 7. O *site* usa Protocolo de Transferência de Hipertexto Seguro (HTTPS)?
- h) 8. O *site* usa uma porta não padrão?
- i) 9. O *site* exibe *pop-ups*?
- j) 10. O *site* tem um *pagerank* menor que 0.2?
- k) 11. O *site* está indexado pelo Google?

Além disso, é preciso que a altura da árvore seja pequena, por isso limitou-se a altura da árvore a 2 e 3. A Tabela 1 apresenta os resultados do treinamento e do desempenho da árvore obtida no conjunto de testes. Pode-se notar que o tempo para

treinamento é muito maior quando a altura é 3, mas apesar disso a solução obtida acerta apenas 0,7% mais classificações no conjunto de treinamento.

Tabela 1 – Treinamento e desempenho da árvore

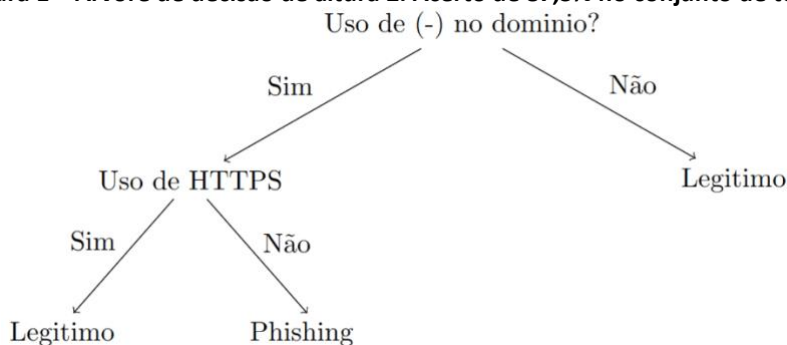
Título	Tempo de treinamento	% acertos treino	% acertos teste
Altura 2	2 segundos	88,0%	87,3%
Altura 3	1507 segundos	88,7%	84,0%

Fonte: Elaboração própria.

Pode-se notar ainda que, no conjunto de teste, a árvore mais simples com altura dois teve um desempenho maior. Conjectura-se que quando a árvore é mais complexa aumenta a possibilidade de acontecer o que é conhecido como *overfitting* (Ying, 2019), quando um sistema se adapta demasiadamente ao conjunto de treinamento, e acaba tendo dificuldade de ser generalizado para outras entradas.

A Figura 1 mostra a árvore que foi obtida quando a altura é limitada a dois, pode-se observar que apenas 2 perguntas simples são capazes de identificar se o *site* é de *Phishing* com uma precisão razoável de 87,3%.

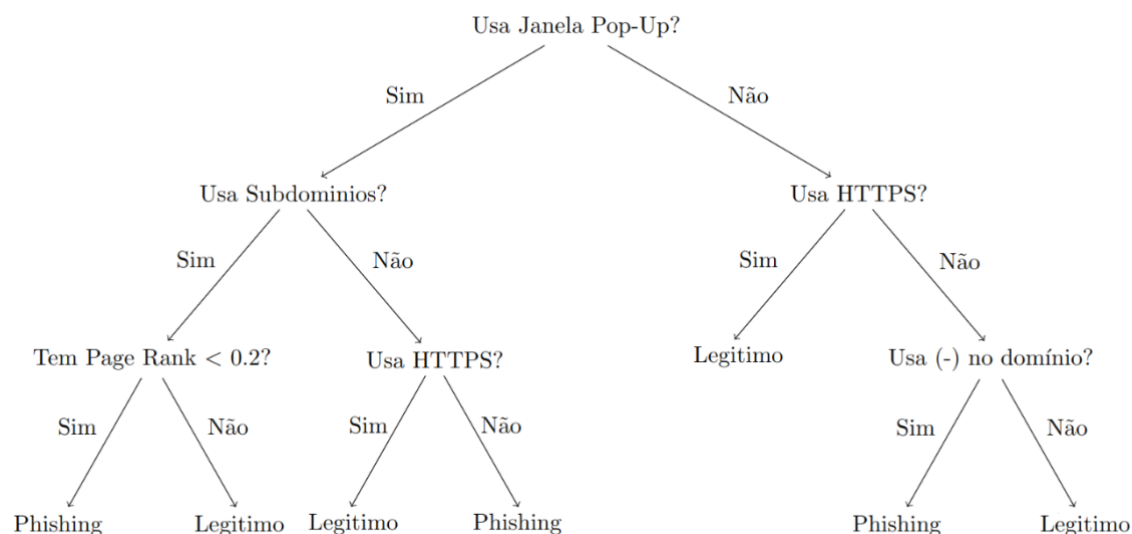
Figura 1 – Árvore de decisão de altura 2. Acerto de 87,3% no conjunto de teste



Fonte: Elaboração própria.

A Figura 2 mostra a árvore que foi obtida quando a altura é limitada a três, pode-se considerar que a complexidade da árvore é bem maior. Ao longo de toda a árvore ocorreram 6 perguntas, sendo duas repetidas. Além disso, a pergunta sobre o PageRank exige uma consulta a alguma ferramenta externa.

Figura 2 – Árvore de decisão de altura 3. Acerto de 84% no conjunto de teste



Fonte: Elaboração própria.

Dessa forma observa-se que nesse caso particular o limite aumentado da altura não trouxe benefícios que o justificassem.

5 CONSIDERAÇÕES FINAIS

Neste trabalho foi apresentada uma abordagem para a detecção de *sites* de *Phishing* utilizando árvores de classificação ótimas. A escolha desse método foi gerar uma instrução de detecção que não dependa da execução de um *software*, ou seja, um usuário pode ser instruído para verificar poucas observações em um *site* e detectar com boa precisão se um *site* se trata de um golpe.

O objetivo geral da pesquisa foi atendido na apresentação da Tabela 1, na qual foi encontrada uma árvore de classificação que atendia os requisitos de ser simples e interpretável por usuários. Já o objetivo específico, foi atendido na Figura 1, na qual elencou-se as orientações mais precisas para a identificação de *sites* de *phishing*.

Uma das principais contribuições deste trabalho reside na demonstração da eficácia de árvores de classificação interpretáveis, construídas através de Programação Linear Inteira (PLI), para a detecção de *sites* de *Phishing*. Foram obtidas árvores de altura 2 e 3, e notavelmente, a árvore com apenas altura 2 alcançou uma precisão de 87,3% no conjunto de testes. Este resultado sublinha a capacidade de um modelo simples, baseado

em apenas duas perguntas sobre características facilmente observáveis de um *site*, de discriminar entre *sites* legítimos e fraudulentos com uma acurácia considerável. Essa simplicidade e interpretabilidade são cruciais, pois permitem que usuários comuns possam ser instruídos a verificar esses poucos elementos em um *site* para se protegerem contra golpes de *phishing*, sem a necessidade de *softwares* especializados.

A pesquisa também observou que uma árvore mais complexa (altura 3) não apresentou uma melhoria significativa na precisão no conjunto de testes, e até mesmo demonstrou sinais de *overfitting*, reforçando o valor de modelos mais simples e interpretáveis para este problema. Em suma, a identificação de uma árvore de decisão rasa e altamente precisa, utilizando um conjunto limitado de atributos facilmente verificáveis, representa um avanço significativo na criação de métodos de detecção de *phishing* acessíveis e compreensíveis para o público em geral

Como trabalhos futuros pretende-se ampliar os testes computacionais para uma melhor validação do método, por exemplo, testar outras métricas como precisão, sensibilidade e pontuação F1, para verificar a robustez do método. Também pode-se utilizar metodologia semelhante para outros tipos de detecção, como notícias e documentos falsos. Além disso, o método também pode ser utilizado para a classificação de outras naturezas, como público-alvo, classificação indicativa (faixa etária), classificação de gênero, entre outras. Pretende-se ainda publicar a implementação dos códigos para a reprodutibilidade da pesquisa.

REFERÊNCIAS

- AGHAEI, Sina; GÓMEZ, Andrés; VAYANOS, Phebe. Strong optimal classification trees. **Operations research**, [S. l.], p. 1-19, 2024. DOI: <https://doi.org/10.1287/opre.2021.0034>. Disponível em: <https://pubsonline.informs.org/doi/10.1287/opre.2021.0034>. Acesso em: 15 maio 2025.
- BERTSIMAS, Dimitris; ORFANOUDAKI, Agni; WIBERG, Holly. Interpretable clustering: an optimization approach. **Machine Learning**, [S. l.], v. 110, p. 89-138, 2021. Disponível em: <https://link.springer.com/article/10.1007/s10994-020-05896-2>. Acesso em: 10 mar. 2025.

CHAND, Eswar. **Phishing website detector**. 2020. Disponível em: <https://www.kaggle.com/datasets/eswarchandt/phishing-website-detector>. Acesso em: 10 mar. 2025.

FIGUEREDO, Jader Duque. **Modelos de aprendizado de máquina para árvore de decisão interpretável: otimização vs heurística**. 2024. Dissertação (Mestrado em Ciência e Tecnologia da Computação) – Universidade Federal de Itajubá, Instituto de Engenharia de Sistemas e Tecnologia da Informação, Itajubá, 2024. Disponível em: <https://repositorio.unifei.edu.br/jspui/handle/123456789/4191>. Acesso em: 10 mar. 2025.

GREENACRE, Michael *et al.* Principal component analysis. **Nature reviews methods primers**, [S. l.], v. 2, n. 100, 2022. Disponível em: <https://www.nature.com/articles/s43586-022-00184-w>. Acesso em: 10 mar. 2025.

KHONJI, Mahmoud; IRAQI, Youssef; JONES, Andrew. Phishing detection: a literature survey. **IEEE communications surveys and tutorials**, [S. l.], v. 15, n. 4, p. 2091-2121, 2013. Disponível em: <http://romisatriawahono.net/lecture/rm/survey/network%20security/Khonji%20-%20Phishing%20Detection%20-%202013.pdf>. Acesso em: 10 mar. 2025.

SAFI, Asadullah; SINGH, Satwinder. A systematic literature review on phishing website detection techniques. **Journal of King Saud University-Computer and Information Sciences**, Saudi Arabia, v. 35, n. 2, p. 590-611, 2023. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1319157823000034>. Acesso em: 15 maio 2025.

SHARIFANI, Koosha; AMINI, Mahyar. Machine Learning and Deep Learning: a review of methods and applications. **World information technology and engineering journal**, [S. l.], v. 10, n. 7, p. 3897-3904, 2023. Disponível em: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4458723. Acesso em: 10 mar. 2025.

VAN DER LINDEN, Jacobus; DE WEERDT, Mathijs; DEMIROVIĆ, Emir. Necessary and sufficient conditions for optimal decision trees using dynamic programming. **Advances in neural information processing systems**, [S. l.], v. 36, p. 9173-9212, 2023. Disponível em: https://proceedings.neurips.cc/paper_files/paper/2023/file/1d5fce9627e15c84db572a66e029b1fc-Paper-Conference.pdf. Acesso em: 15 maio 2025.

WOLSEY, Laurence Alexander; NEMHAUSER, George Lann. **Integer and combinatorial optimization**. [Hoboken]: Wiley, 1999.

WOLSEY, Laurence Alexander. **Integer programming**. Hoboken: Wiley, 2020.

YING, Xue. An overview of overfitting and its solutions. **Journal of Physics: Conf. Series**, [S. l.], v. 1168, n. 2, e022022, 2019. Disponível em:

<https://iopscience.iop.org/article/10.1088/1742-6596/1168/2/022022/pdf>. Acesso em: 10 mar. 2025.



[4.0 Internacional](https://creativecommons.org/licenses/by/4.0/)

Esta licença permite compartilhamento, remixe, adaptação e criação a partir do trabalho, mesmo para fins comerciais, desde que sejam atribuídos créditos ao(s) autor(es). Conteúdos elaborados por terceiros, citados e referenciados nesta obra não são cobertos pela licença.

Como citar este trabalho:

HOKAMA, Pedro Henrique Del Bianco; FIQUEREDO, Jader Duque; CAVALHEIRO, Karen Cristina Soares. *Aprendizado de Máquina para detecção de sites de Phishing*. In: WORKSHOP DE INFORMAÇÃO DADOS E TECNOLOGIA, 8., 2025, Marília, SP. **Anais [...]**. Marília, SP: Universidade de Marília, 2025. DOI: <https://doi.org/10.22477/viii.widat.271>.



WIDaT 2025

VIII WORKSHOP DE INFORMAÇÃO,
DADOS E TECNOLOGIA

MARÍLIA • SP

Reflexões acerca da Intencionalidade Informacional e da Inteligência Artificial Generativa no contexto acadêmico-científico: explorando o uso dos *chatbots*

*Reflections on Informational Intentionality and Generative Artificial Intelligence in the
academic-scientific context: exploring the use of chatbots*

Fernanda Alves Sanchez¹



<http://lattes.cnpq.br/5293590576574461>



<https://orcid.org/0000-0003-1543-2773>

Eduardo Guimarães Mielo²



<http://lattes.cnpq.br/1486758964403896>



<https://orcid.org/0000-0003-2305-3706>

Silvana Ap. B. G. Vidotti³



<http://lattes.cnpq.br/7390573927636069>



<https://orcid.org/0000-0002-4216-0374>

Resumo

O uso de ferramentas de Inteligência Artificial (IA) se mostra cada vez mais frequente no meio acadêmico-científico. Este estudo explora as relações conceituais entre a Intencionalidade Informacional e a IA Generativa, com foco no uso de *chatbots*. A pesquisa tem como objetivo geral analisar o desempenho da ferramenta *DeepSeek* no contexto acadêmico-científico e estabelecer reflexões conceituais entre a Intencionalidade Informacional e a IA Generativa. A metodologia adotada é de natureza exploratória, qualitativa e descritiva, utilizando observação direta e aplicação prática do *DeepSeek*. Os resultados indicam que o *DeepSeek* pode otimizar o tempo de pesquisa, especialmente em processos de filtragem e organização de dados, mas também destacam desafios relacionados à dependência excessiva dessas ferramentas, que podem comprometer o aprendizado humano e a qualidade das informações geradas. Como considerações finais reforça-se a necessidade de um uso consciente e complementar da IA Generativa, sem substituir a Intencionalidade humana. Por fim, destaca-se que, embora as ferramentas de IA Generativa sejam úteis, é essencial manter um olhar crítico e ético sobre seu uso no ambiente acadêmico-científico.

Palavras-chave: intencionalidade informacional; inteligência artificial generativa; DeepSeek.

¹ Doutorado em Ciência da Informação, Universidade Estadual Paulista Júlio de Mesquita Filho (Unesp), Marília, São Paulo, Brasil. E-mail: fernanda.a.sanchez@unesp.br.

² Graduação em Pedagogia, Universidade Estadual Paulista Júlio de Mesquita Filho (Unesp), Marília, São Paulo, Brasil. E-mail: eduardo.mielo@unesp.br.

³ Doutorado em Educação, Universidade Estadual Paulista Júlio de Mesquita Filho (Unesp), Marília, São Paulo, Brasil. Docente no Departamento de Graduação e Pós-Graduação em Ciência da Informação da Universidade Estadual Paulista Júlio de Mesquita Filho (Unesp), Marília, São Paulo, Brasil e Coordenadora da Coordenadoria de Tecnologias Aplicadas (COTEA) do Ibict, Brasília, Distrito Federal, Brasil. E-mail: silvana.vidotti@unesp.br.

Abstract

The use of Artificial Intelligence (AI) tools is becoming more frequently used in the academic-scientific environment. This study explores the conceptual relationships between Informational Intentionality and Generative AI, focusing on the use of chatbots. The research has the objective of analyzing the performance of the DeepSeek tool in the academic-scientific context and establishing conceptual reflections between Informational Intentionality and Generative AI. The methodology adopted is exploratory, qualitative and descriptive, using direct observation and practical application of DeepSeek. The results indicate that DeepSeek can optimize search time, especially in data filtering and organization processes, but it is also important to highlight challenges related to excessive dependence on these tools, which can compromise human learning and the quality of the information generated. As final considerations, the need for a conscious and complementary use of Generative AI is reinforced, without replacing human Intentionality. Finally, we conclude that, although Generative AI tools are useful, it is essential to maintain a critical and ethical look at their use in the academic-scientific environment.

Keywords: Informational Intentionality; Generative Artificial Intelligence; DeepSeek.

1 INTRODUÇÃO

No âmbito das Tecnologias de Informação e Comunicação (TIC), destacam-se as inovações decorrentes de técnicas e tecnologias ligadas à Inteligência Artificial (IA), como por exemplo, o Processamento de Linguagem Natural (PLN) e o *Machine Learning*, que são consideradas aliadas na melhoria do aprendizado de máquina, o que acarreta em previsões, interpretações e resolução de problemas com certa agilidade.

Dentre os tipos de Inteligência Artificial, a IA Generativa tem ganhado cada vez mais destaque, principalmente após o surgimento do ChatGPT (*Generative Pretrained Transformer*). Esse tipo de IA desenvolve ferramentas tecnológicas conhecidas como *chatbots* capazes de “[...] gerar textos, imagens, áudios e os mais diversos tipos de conteúdo, com base em treinos realizados com conjuntos de dados existentes, dependendo do domínio de aplicação” (Coneglian *et al.*, 2023, p. 2).

Nota-se neste cenário uma relação entre os aspectos humanos e tecnológicos, tendo em vista a troca entre as características humanas e o aprendizado automatizado da máquina. Essa relação dentro do contexto da Ciência da Informação pode ser vista na concepção de Intencionalidade Informacional, a qual se destaca, a junção da Intencionalidade Humana e a Intencionalidade Tecnológica, por meio das mediações. Entende-se as mediações como interferências construtivas, uma vez que as tecnologias moldam as ações humanas, e os humanos, por sua vez, moldam as funcionalidades das tecnologias (Sanchez, 2024).

Huang e Tan (2023), bem como a Nature (2023) destacam o crescente uso de ferramentas tecnológicas, como os *chatbots*, no contexto acadêmico-científico. Os autores

também destacam os desafios emergentes nas interações humanas e tecnológicas, especialmente no que diz respeito às implicações éticas dessas relações.

O uso de bases de dados nacionais e internacionais é uma etapa essencial no processo de pesquisa para qualquer pesquisador, para realizar o levantamento bibliográfico. Contudo, em determinadas áreas temáticas, o volume expressivo de resultados apresentados por essas bases pode representar um desafio significativo, exigindo critérios rigorosos de seleção e análise por parte do pesquisador.

A problemática deste estudo refere-se às reflexões sobre a concepção da Intencionalidade Informacional, um conceito emergente no campo da Ciência da Informação, bem como aos impactos que as ferramentas de IA Generativa podem ocasionar no contexto acadêmico-científico, especialmente no desenvolvimento de pesquisas, a partir de uma análise realizada no *DeepSeek*⁴.

O objetivo geral desta pesquisa é analisar o desempenho da ferramenta *DeepSeek* no contexto acadêmico-científico e promover reflexões conceituais sobre a relação entre a Intencionalidade Informacional e a IA Generativa.

A relevância deste estudo reside no fato de que a Inteligência Artificial e suas ferramentas tecnológicas são uma realidade no mundo contemporâneo, ultrapassando o campo das idealizações. Dessa forma, se faz necessário promover reflexões acerca de aspectos sociais, éticos, técnicos e tecnológicos do uso dessas ferramentas para que seus impactos e interferências sejam positivos para a sociedade. No contexto acadêmico-científico, destaca-se a importância de um uso consciente dessas tecnologias, visando propiciar a colaboração e a comunicação científica.

1.1 IA Generativa e Ferramentas de Chatbots

Nas últimas décadas, a integração generalizada de tecnologias de Inteligência Artificial impulsionou mudanças transformadoras em diversas áreas, abrangendo desde operações comerciais e serviços de saúde até práticas educativas e investigação científica. Desta forma, a utilização da IA, redefiniu paradigmas tradicionais, criando novas oportunidades de inovação e eficiência em distintos contextos.

⁴ DEEPSEEK. **DeepSeek**. Versão V3. c2025. Disponível em: <https://www.deepseek.com/>. Acesso em: 20 fev. 2025.

Diversas ferramentas de IA, especialmente voltadas para o processamento de texto, têm sido desenvolvidas e utilizadas com frequência crescente no meio acadêmico. No entanto, as diretrizes para a integração do conteúdo gerado pela IA com a contribuição humana ainda estão em evolução (Razack *et al.*, 2021). Embora tais ferramentas possam aumentar a eficiência, economizar tempo e apoiar tarefas como planejamento, estudo e redação, elas também representam riscos para a integridade acadêmica (Okaibedi, 2023).

No que diz respeito à perspectiva tecnológica, o anúncio da criação e ampla disponibilização da ferramenta *ChatGPT* para testes, no final de 2022, representou um marco no processo de incorporação de tecnologias na produção acadêmica, da mesma forma que o Google representou um salto na forma como os indivíduos passaram a buscar informações sobre os mais variados temas, inclusive no meio acadêmico (Tatnall; Davey, 2014).

Com uma interface amigável em formato de *chat*, fácil navegação e disponibilidade ampla e irrestrita – sendo necessário apenas o fornecimento de um endereço de e-mail no site do desenvolvedor-, o *ChatGPT* apresenta potencial para se tornar mais uma ferramenta tecnológica relevante no ambiente acadêmico global (Peres, 2024). No entanto, sua sofisticação tecnológica também representa riscos, uma vez que pode ser utilizada como atalho eticamente questionável no processo de construção e disseminação do conhecimento acadêmico.

Nesse contexto, como escolha de ferramenta para análise apresenta-se o *DeepSeek*, recentemente lançado no dia 20 de janeiro de 2025 que causou impacto mundial relevante em escala global em um curto espaço de tempo. Além disso, a análise com um *chatbot* novo no mercado e pouco analisado na comunidade acadêmica agregou relevância à pesquisa. Segundo o pesquisador Cleber Zanchettin na Universidade Federal de Pernambuco (UFPE), em entrevista para o BBC, “além da alta tensão na arena dos negócios e da geopolítica, a inovação em si trazida pela plataforma impressionou a comunidade científica” (Mota, 2025).

Isso porque, a ferramenta trouxe diferenciais aos que pesquisam ou empreendem com esse tipo de IA Generativa. De acordo com o pesquisador, o *DeepSeek* diferente de outras ferramentas apresenta alguns avanços tecnológicos significativos, como o código aberto, o raciocínio explícito, a aprendizagem por reforço, entre outras técnicas e modelos inovadores, configurando um conjunto de recursos que não havia em outras plataformas até o momento (Mota, 2025).

1.2 Intencionalidade Informacional

Na Ciência da Informação, o conceito de Intencionalidade ainda é uma temática emergente e pouco abordada. Trata-se de um conceito advindo dos estudos teóricos do campo da Filosofia, nas escolas de pensamento da Fenomenologia e da Pós-Fenomenologia, em áreas da Filosofia da Mente e da Filosofia da Tecnologia.

Os estudos sobre a Intencionalidade na Fenomenologia têm relação com a consciência e a direcionalidade das ações dos sujeitos informacionais, que podem ser materializadas por meio das representações, seja nos atos de fala (linguagem) ou pela percepção (expressões visuais) (Searle, 1993).

A Intencionalidade na Pós-Fenomenologia é denominada como Intencionalidade Tecnológica. Trata-se de um conceito que visa adaptar a ciência as mudanças do século XXI visando a interação entre o ser humano e a tecnologia, ampliando o que, na Fenomenologia, é visto como a relação entre o homem-mundo. Na Pós-Fenomenologia, é entendida como uma interligação entre o homem-tecnologia-mundo (Ihde, 1990).

No campo da CI, Sanchez (2024) apresenta a concepção da Intencionalidade Informacional, que traz como nexos a mediação entre os conceitos da Intencionalidade Humana (IH) e da Intencionalidade Tecnológica (IT), destacando a existência de abordagens teóricas, técnicas e tecnológicas que podem ser classificadas e utilizadas sob os aspectos de interferência mútua entre os aspectos humanos e tecnológicos em todos os processos informacionais.

A IH refere-se à construção da bagagem cognitiva dos sujeitos informacionais que ocorre por meio de toda e qualquer experiência vivida ao longo da vida em seus diferentes contextos – sociais, culturais, políticos, etc. – passa a adquirir novas habilidades, competências, comportamentos, sentimentos, crenças etc., tornando-o um ser humano único. Ou seja, há a presença de subjetividades e singularidades que apenas os seres humanos são capazes de demonstrar (Sanchez, 2024). Por outro lado, na IT, o foco está na materialização, ou seja, nas tecnologias (*hardware* e/ou *software*). Ela “[...] apresenta a autonomia das tecnologias, indo além da ação de mediação nas experiências humanas” (Sanchez, 2024, p. 52).

2 PROCEDIMENTOS METODOLÓGICOS

Trata-se de uma pesquisa exploratória, qualitativa, descritiva e aplicada sobre o uso da ferramenta *chatbot* no contexto acadêmico-científico, com foco na análise da ferramenta *DeepSeek*. A partir de observação direta dos pesquisadores e aplicação prática, foi possível analisar o desempenho da ferramenta e elaborar reflexões iniciais acerca da relação conceitual entre a Intencionalidade Informacional e o uso das ferramentas de IA Generativa no âmbito-acadêmico científico.

Dessa forma, a pesquisa aplicada foi realizada em 4 etapas. A primeira etapa foi a realização do levantamento bibliográfico nas bases de dados Portal Brasileiro de Publicações e Dados Científicos em Acesso Aberto (Oasis) e na Biblioteca Digital Brasileira de Teses e Dissertações (BDTD) do Ibict, a partir da utilização dos descritores “*accountability*”, “*responsabilização educacional*”, “*São Paulo*” e “*paulista*”, a qual foram encontrados uma grande quantidade de resultados. As plataformas possibilitam ao usuário fazer o *download* dos resultados gerados em uma planilha no formato CSV, totalizando 16 arquivos.

Vale ressaltar que a temática do levantamento se justifica pela relação com a dissertação de mestrado de um dos autores deste trabalho, que discute o processo de responsabilização educacional (*accountability*) – ou seja, a premiação, sanção e bonificação de escolas a partir do resultado das avaliações de larga escala – nas redes estaduais de ensino.

Quadro 1 – Levantamento bibliográfico⁵

Descritores	BDTD	Oasis	Descritores	BDTD	Oasis
Responsabilização (T) + São Paulo (C)	52	172	Responsabilização (C) + São Paulo (C) + Educação (C)	154	277
Responsabilização (T) + Paulista (C)	7	16	Responsabilização (C) + Paulista (C) + Educação (C)	48	100
Accountability (T) + São Paulo (C)	50	144	Accountability (C) + São Paulo (C) + Educação (C)	163	259
Accountability (T) + Paulista (C)	11	23	Accountability (C) + Paulista (C) + Educação (C)	52	84

Fonte: Elaboração própria.

⁵ Legenda: C – Campo; T – Título

Diante de um grande volume de dados, é necessário eliminar itens duplicados e remover artigos que se distanciam do objeto de estudo da pesquisa. Para otimizar esse processo utilizou-se o *DeepSeek* que oferece três opções de refinamento para remoção de duplicatas: manual, semi-automatizado e automatizado.

No processo manual, o usuário realiza a análise individual de cada item, o que demanda um tempo maior de análise de dados. No processo semi-automatizado, o *DeepSeek* instrui o usuário a consolidar manualmente todas as planilhas em um único arquivo, e em seguida utilizar as funcionalidades do software de planilhas para remover as duplicatas. Por fim, o processo automatizado é realizado por meio de um código gerado pela própria IA em *Python*⁶, dispensando a necessidade de junção manual dos arquivos.

Na etapa seguinte, foi necessário remover os dados de trabalhos que se distanciavam do escopo da temática da pesquisa. O arquivo gerado por meio da programação supramencionada foi submetido ao *chatbot* para a realização de filtragem. Desta forma, usou-se de prompts para refinar os dados a partir dos resumos. Solicitou-se ao *chatbot* que indicasse as linhas do arquivo que não continham os descritores utilizados no levantamento bibliográfico, dispensando a repetição dos resumos nesta indicação. Com essas informações, a exclusão dos registros foi realizada de forma manual na planilha. Por fim, o pesquisador ainda removeu oito linhas de dados por delimitação da pesquisa.

Na próxima seção, serão apresentados os resultados referentes de uma análise comparativa com outras ferramentas de *chatbots*, bem como discussões conceituais relacionadas ao conceito de Intencionalidade Informacional.

3 RESULTADOS

Em um contexto geral, o principal propósito de utilizar ferramentas de IA generativas é, preferencialmente, a redução de tempo necessário para a execução de uma tarefa. Isso pode ser benéfico, por exemplo, em situações como a combinação do curto período de duração dos programas de pós-graduação (curso de mestrado) e o extenso volume de

⁶ Python é uma linguagem de programação de alto nível, interpretada de script, imperativa, orientada a objetos, funcional, de tipagem dinâmica e forte (Python Software Foundation, c2025).

resultados obtidos em temáticas que se interrelacionam em diferentes áreas do conhecimento.

Trindade e Oliveira (2024), com base em diversos autores, compilaram algumas possibilidades de aplicação das IA Generativas no contexto acadêmico-científico, conforme Quadro 2.

Quadro 2 – Possibilidades no uso da IA Generativa no contexto acadêmico-científico

Possibilidades no uso da IA Generativa no contexto acadêmico-científico	
1	Gerenciar, analisar e resumir grandes quantidades de informações e dados;
2	Coletar dados amostrais e gerar cenários;
3	Encontrar artigos acadêmicos;
4	Obter o resumo de um estudo consultado, o que permite consultar informações/dados com maior celeridade e compreender rapidamente os pontos principais de um texto, organizar;
5	Elaborar o resumo da sua pesquisa;
6	Melhorar/treinar a escrita (reescrever uma frase particularmente complexa de maneira mais clara, por exemplo);
7	Realizar traduções e/ou escrever textos em outros idiomas;
8	Gerar um rascunho inicial de um artigo científico; i) gerar um título para uma publicação;
9	Gerar um título para uma publicação;
10	Gerar recursos visuais (figuras, tabelas, infográficos e outros) que auxiliam na síntese, na clareza e na compreensão de dados;
11	Planejar aulas (criar planos de aula, atividades, questionários, problemas práticos, etc.).

Fonte: Trindade e Oliveira (2024).

Nota-se que os benefícios do uso das ferramentas de IA Generativa no processo de pesquisa vão além da simples redução de tempo. Em determinadas situações, a tecnologia passa a ser uma “parceira”, na formulação de novas hipóteses ou, até mesmo, na ampliação das perspectivas sobre questões específicas, como por exemplo, a elaboração de um título para publicação acadêmica.

Mielo *et al.* (2024), em parceria com o *Institute of Czech-Brazilian Academic Cooperation* (INCBAC) e pesquisadores do Brasil e da República Tcheca, realizaram uma pesquisa baseada em testes em diversas ferramentas gratuitas de IA Generativa, tais como: *ChatPDF*, *DeepL Write*, *EinBlick*, *Explain Paper*, *Jenni*, *Open Read*, *Scholarcy* e *Scispace*. O estudo abordou pontos fortes e fracos dessas ferramentas no âmbito das pesquisas em humanidades, especialmente no contexto de resumo e redação de textos acadêmicos.

Os autores concluíram que

[...] as ferramentas de IA economizam tempo, são fáceis de usar e apresentam resultados confiáveis [...] são uma oportunidade para complementar estudos, encontrar *insights* e entender melhor relações conceituais (Mielo *et al.*, 2024, p. 5, tradução nossa)⁷.

Entretanto, os pesquisadores também destacaram pontos críticos relevantes para o contexto acadêmico-científico, em especial no que diz respeito à ameaça ao processo de aprendizagem. Alertam para o risco do uso exclusivo dessas ferramentas para substituir a tradicional leitura dos artigos. Isso porque, a substituição pode afetar negativamente, ao longo do tempo, o processo de aprender, ou seja, afetar diretamente a plasticidade cerebral⁸ do ser humano.

Além disso, a pesquisa destaca problemas no que se refere às versões gratuitas dessas ferramentas, indicando limitações em seu uso. Dessa forma, os autores indicam, para certas situações, a utilização de mais de uma ferramenta para a elaboração de textos, visando de certa forma enriquecer e validar as informações e o conteúdo gerado.

Diante do exposto, a proposta deste presente estudo contribui para a análise de uma nova ferramenta de IA generativa, com o objetivo de verificar sua eficiência, embora com um enfoque mais quantitativo do que qualitativo no que se refere à redação de textos. Assim, a escolha pelo uso do *DeepSeek* visou otimizar o tempo gasto com o levantamento bibliográfico de uma temática de pesquisa extensa, no caso, a *Accountability*. Considerando que, em comparação à análise manual, houve uma diferença de oito que foram retirados pelo pesquisador com base nos critérios de recorte da pesquisa, conclui-se que a ferramenta atendeu a essa necessidade de otimização do tempo.

Entre as principais diferenças do *ChatGPT* e *DeepSeek*, destaca-se o fato de que o *ChatGPT* não possui código aberto, ou seja, seus modelos de códigos não são completamente acessíveis para modificação ou uso irrestrito. Em contrapartida, o *DeepSeek* promove uma

⁷ Original: “[...] that AI tools are timesaving, easy to use, and present reliable results. [...] e an opportunity to complement studies, find insights, and better understand conceptual relationships” (Mielo *et al.*, 2024, p. 5).

⁸ A plasticidade cerebral refere-se à “[...] noção de capacidade adaptativa do cérebro em diferentes espectros [...]” (Costa; Silva; Jacobsen, 2019, p. 460). A plasticidade é algo que irá perpetuar ao longo da vida e que ocorrerá por conta dos diferentes estímulos de aprendizagem nos diversos contextos da existência do sujeito no mundo.

abordagem de código aberto, permitindo que desenvolvedores e pesquisadores acessem e modifiquem o modelo, propiciando a colaboração e a inovação entre os usuários.

Além disso, vale ressaltar que, assim como na pesquisa de Mielo *et al.* (2024) que sugere o uso combinado de diferentes ferramentas de IA Generativa para aprimorar resumos e redações, esta pesquisa também entende que a pesquisa qualitativa demanda a combinação de várias ferramentas, por exemplo, os *chatbots* ou *softwares* que apoiam a criação de planilhas, para enriquecimento do processo de análise.

Trazendo essa perspectiva para as reflexões acerca da temática da Intencionalidade Informacional, é importante ressaltar dois pontos. O primeiro é que sempre haverá a consciência humana por trás dos resultados obtidos pela máquina. Comandos bem elaborados fazem com que os resultados sejam melhores, evidenciando que a Intencionalidade dos sujeitos direciona o desempenho da máquina e vice-versa, e isso tem muita relação com as discussões sobre aprendizagem mencionadas anteriormente.

Além disso, como um segundo ponto Sanchez (2024) defende a ideia de que trabalhos com equipes multidisciplinares enriquecem o projeto/pesquisa, pois são olhares diferentes sobre um mesmo assunto, evitando os vieses que os sujeitos podem carregar em determinadas ações (intencionalidade humana), assim, de acordo com as recomendações do uso de diferentes ferramentas tecnológicas nota-se que a intencionalidade tecnológica também está direcionada para as ações em conjunto para um melhor desempenho.

No que se refere ao uso das ferramentas de IA Generativa, entende-se que é possível utilizá-las no auxílio do trabalho de pesquisa, mas cabe constantemente o olhar crítico e atento do pesquisador aos resultados. É fato de que elas podem ser úteis para sistematizar a abordagem manual, no entanto ao considerar que esses tipos de ferramentas oferecem recurso de bate-papo para interagir e obter respostas, é crucial entender — e praticar — os tipos de *prompts*⁹ usados, especialmente com versões gratuitas.

Para otimizar os resultados dessas ferramentas, recomenda-se a utilização *prompts* objetivos e bem direcionados, para que as respostas também se apresentem da mesma forma. Por exemplo, pergunta como "Qual é a metodologia para o resultado?" pode gerar uma

⁹ Termo utilizado para um comando ou solicitação em sistemas informacionais que geralmente utilizam modelo de linguagem.

explicação superficial daquela seção do artigo. Por outro lado, um *prompt* mais específico como "Por que essa metodologia foi usada neste artigo?" tende fornecer uma análise mais detalhada e assertiva do contexto.

Dando continuidade às reflexões sobre a Intencionalidade Informacional, é importante considerar aspectos sociais, éticos e geopolíticos, como evidenciado pelo lançamento do *DeepSeek*, que trouxe inovações no âmbito da IA Generativa e conseguiu abalar o mercado e a bolsa de valores de empresas renomadas mundialmente.

Entende-se que a Intencionalidade Informacional, nesse lançamento, sob a perspectiva da Intencionalidade humana, pode ser vista tanto de forma positiva quanto negativa. Na primeira, destaca-se o fato de apresentar ao mundo uma possibilidade de igualar os modelos de ferramentas de *chatbots*, permitindo que mais empresas possam desenvolver e competir de forma mais igualitária no mercado de trabalho. Entretanto, ainda falando sobre a Intencionalidade humana, como ponto negativo, pode-se perceber um grande embate geopolítico, além de algumas relações de ego e poder.

No que se refere à Intencionalidade Tecnológica do *DeepSeek*, a ferramenta apresenta um recurso considerado como uma melhoria em seu desempenho que é o "afastamento do humano", conhecido tecnicamente como *Human-In-The-Loop* (HITL, na sigla em inglês), que é mais utilizado no refinamento da ferramenta quanto aos resultados obtidos (Mota, 2025).

No caso do *DeepSeek*, esse afastamento reduz significativamente o custo operacional da ferramenta, pois ela adota um modelo de aprendizado por reforço, sendo treinada a partir de recompensas dentro do próprio sistema (Mota, 2025). No entanto, com o afastamento do humano, a qualidade das respostas corretas pode ser comprometida e pensando em aspectos sociais e geracionais, onde estudantes de todas as idades utilizam esse tipo de ferramenta, os conhecimentos adquiridos e que formalizam a Intencionalidade — incluindo as competências (aprendizagem), crenças e emoções dos sujeitos — podem sofrer graves impactos devido a esse afastamento.

4 CONSIDERAÇÕES FINAIS

Retomando o objetivo geral desta pesquisa, que visa analisar o desempenho da ferramenta *DeepSeek* no contexto acadêmico-científico e promover reflexões conceituais

sobre a relação entre a Intencionalidade Informacional e a IA Generativa, é possível afirmar que o objetivo foi alcançado. Isso se comprova pelas discussões acerca do desempenho da ferramenta, seus pontos positivos e negativos, e suas relações com a concepção da Intencionalidade Informacional, por meio das mediações humanas e tecnológicas presentes nos processos de desenvolvimento de pesquisa no âmbito acadêmico-científico.

Fica claro, a partir deste estudo e dos demais trabalhos apresentados ao longo dos resultados, que as ferramentas de IA Generativa são uma realidade tanto no contexto acadêmico-científico quanto em outros setores da sociedade. Tais ferramentas podem contribuir em diferentes etapas dos processos de pesquisa. Contudo, destaca-se que elas não devem substituir as ações humanas, mas atuar de forma complementar, estabelecendo relações mútuas de mediação humanas e tecnológicas.

As reflexões realizadas neste texto também permitem identificar novas hipóteses de pesquisa relacionadas à concepção da Intencionalidade Informacional, especialmente no que se refere à realização de estudos que busquem compreender a Intencionalidade maquínica estabelecida na pós-mediação entre a Intencionalidade humana e tecnológica. Em outras palavras, trata-se de investigar o processo que ocorre após o desenvolvimento de *hardware* e/ou *software* por seres humanos e a transição para o aprendizado autônomo da máquina, onde ela aprende a partir de seus próprios dados.

Por fim, destaca-se a necessidade de uma atenção constante para que o uso das ferramentas de IA Generativa não se torne uma ameaça ao aprendizado e ao desenvolvimento cognitivo dos sujeitos informacionais, mas sim um aliado na colaboração científica e na busca por soluções para as problemáticas da sociedade. Para tanto, torna-se essencial investir em ações e formações continuadas para professores e estudantes, de modo a preparar a população para o uso consciente e ético dessas ferramentas. Isso se justifica pela própria acessibilidade da ferramenta. Como apresentado no decorrer do texto, o compromisso ético dos pesquisadores deve estar orientado para compreender e utilizar essa ferramenta de forma instruída ao invés de tentar impedir o uso de tecnologias que, cada vez mais, se fazem presente no cotidiano humano.

REFERÊNCIAS

- CONEGLIAN, C. S.; TORINO, E.; SANTARÉM SEGUNDO, J. E.; VIDOTTI, S. A. B. G. Inteligência artificial generativa e recuperação da informação: tendências e oportunidades de pesquisa. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 23., 2023, Aracaju. **Anais [...]**. Aracaju: ENANCIB, 2023. Disponível em: <https://ancib.org/enancib/index.php/enancib/xxxiiienancib/paper/viewFile/1944/1401>. Acesso em: 20 fev. 2025.
- COSTA, A. R.; SILVA, P. L. O.; JACÓBSEN, R. T. Plasticidade cerebral: conceito(s), contribuições ao avanço científico e estudos brasileiros na área de Letras. **Entrepalavras**, Fortaleza, v. 9, n. 3, p. 457-476, set-dez/2019. DOI: <http://dx.doi.org/10.22168/2237-6321-31445>. Disponível em: <http://www.entrepalavras.ufc.br/revista/index.php/Revista/article/view/1445>. Acesso em: 20 fev. 2025.
- HUANG, J.; TAN, M. The role of ChatGPT in scientific communication: writing better scientific review articles. **American journal of cancer research**, [S. l.], v. 13, n. 4, p. 1148-1154, apr. 2023. Disponível em: <https://pmc.ncbi.nlm.nih.gov/articles/PMC10164801/>. Acesso em: 20 fev. 2025.
- IHDE, D. **Technology and the lifeworld**: from garden to earth. Bloomington: Indiana University Press, 1990.
- MIELO, E. G.; COSTA, N. V.; LIMA, R. S.; DOČEKAL, V. **Effectiveness of humanities research using AI**: testing the tools for text processing. [Praga]: INCBAC INSTITUTE, 2024. Disponível em: <https://www.incbac.org/wp-content/uploads/2024/05/UNIGOU-Remote-Publication-2024-Eduardo-Guimaraes-Mielo.pdf>. Acesso em: 20 fev. 2025.
- MOTA, C. V. Pesquisador brasileiro em IA explica por que DeepSeek impressionou: “fizeram de forma totalmente diferente da maioria das empresas de tecnologia”. **BBC NEWS Brasil**, 31 jan. 2025. Disponível em: <https://www.bbc.com/portuguese/articles/cde96995dryo>. Acesso em: 20 fev. 2025.
- NATURE. Tools such as ChatGPT threaten transparent science; here are our ground rules for their use. **Nature**, [S. l.], v. 613, n. 7945, p. 612, jan. 2023. DOI: <http://doi.org/10.1038/d41586-023-00191-1>. Disponível em: <https://doi.org/10.1038/d41586-023-00191-1>. Acesso em: 20 fev. 2025.
- OKAIBEDI, D. E. ChatGPT and the rise of generative AI: threat to academic integrity? **Journal of Responsible Technology**, [S. l.], v. 13, p. 1-4, 100060, apr. 2023. Disponível em: <https://doi.org/10.1016/j.jrt.2023.100060>. Acesso em: 20 fev. 2025.
- PERES, F. A literacia em saúde no ChatGPT: explorando o potencial de uso de inteligência artificial para a elaboração de textos acadêmicos. **Ciência e Saúde Coletiva**, [S. l.], v. 29, n. 1, e02412023, 2024. DOI 10.1590/1413-81232024291.02412023. Disponível em: <https://doi.org/10.1590/1413-81232024291.02412023>. Acesso em: 20 fev. 2025.

PYTHON SOFTWARE FOUNDATION. **About Python**. c2025. Disponível em: <https://www.python.org/about/>. Acesso em: 20 fev. 2025.

RAZACK, H. I. A.; MATHEW, S. T.; SAAD, F. F. A.; ALQAHTANI, S.A. Artificial intelligence-assisted tools for redefining the communication landscape of the scholarly world. **Science Editing**, [Seoul], v. 8, n. 2, p. 134-144, 2021. Disponível em: <https://doi.org/10.6087/kcse.244>. Acesso em: 20 fev. 2025.

SANCHEZ, F. A. **Intencionalidade informacional**: abordagens teóricas, técnicas e tecnológicas no campo da ciência da informação. 2024. Tese (Doutorado em Ciência da Informação) — Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2024. Disponível em: <https://hdl.handle.net/11449/256900>. Acesso em: 20 fev. 2025.

SEARLE, J. R. **Intentionality**: an essay in the philosophy of mind. Cambridge: Cambridge University Press, 1983.

TATNALL, A.; DAVEY, B. (ed.). **Reflections on the history of computers in education**: early use of computers and teaching about computing in schools. [Heidelberg]: Springer, 2014.

TRINDADE, A. S. C. E.; OLIVEIRA, H. P. C. Inteligência artificial (IA) generativa e competência em informação: habilidades informacionais necessárias ao uso de ferramentas de IA generativa em demandas informacionais de natureza acadêmica-científica. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 29, e-47485, 2024. DOI: <http://doi.org/10.1590/1981-5344/47485>. Disponível em: <https://doi.org/10.1590/1981-5344/47485>. Acesso em: 20 fev. 2025.



[4.0 Internacional](https://creativecommons.org/licenses/by/4.0/)

Esta licença permite compartilhamento, remixe, adaptação e criação a partir do trabalho, mesmo para fins comerciais, desde que sejam atribuídos créditos ao(s) autor(es). Conteúdos elaborados por terceiros, citados e referenciados nesta obra não são cobertos pela licença.

COMO CITAR ESTE TRABALHO:

SANCHEZ, Fernanda Alves; MIELO, Eduardo Guimarães; VIDOTTI, Silvana Aparecida Borsetti Gregorio. Reflexões acerca da Intencionalidade Informacional e da Inteligência Artificial Generativa no contexto acadêmico-científico: explorando o uso dos chatbots. In: WORKSHOP DE INFORMAÇÃO DADOS E TECNOLOGIA, 8., 2025, Marília, SP. **Anais [...]**. Marília, SP: Universidade de Marília, 2025. DOI: <https://doi.org/10.22477/viii.widat.262>.



WIDaT 2025

VIII WORKSHOP DE INFORMAÇÃO,
DADOS E TECNOLOGIA

MARÍLIA • SP

Aplicação da Inteligência Artificial Generativa na Experiência do Usuário: potencialidades e desafios

Application of Generative Artificial Intelligence in User Experience: potentials and challenges

Gustavo Camossi¹



<http://lattes.cnpq.br/2873515398619762>



<https://orcid.org/0000-0002-1553-1053>

Ananda Fernanda de Jesus²



<http://lattes.cnpq.br/5562746387565465>



<https://orcid.org/0000-0001-7873-6040>

Suellen Timm Barros³



<http://lattes.cnpq.br/8245514969497131>



<https://orcid.org/0000-0001-9124-0242>

José Eduardo Santarem Segundo⁴



<http://lattes.cnpq.br/5562746387565465>



<https://orcid.org/0000-0002-0702-7586>

Silvana Aparecida Borsetti Gregorio Vidotti⁵



<http://lattes.cnpq.br/7390573927636069>



<https://orcid.org/0000-0002-4216-0374>

Resumo

A Inteligência Artificial Generativa tem se destacado no campo da Experiência do Usuário ao automatizar processos, personalizar interfaces e otimizar testes. No entanto, sua aplicação também levanta questões sobre ética e a colaboração entre humanos e máquinas. Este estudo tem como objetivo identificar como a Inteligência Artificial Generativa tem sido discutida e aplicada na Experiência do Usuário, destacando seus desafios e potenciais aplicações. Trata-se de uma pesquisa de abordagem exploratória e descritiva, com foco qualitativo,

¹ Doutorando em Ciência da Informação, Universidade Estadual Paulista Júlio de Mesquita Filho (Unesp), Marília, SP, Brasil. Docente no Centro Universitário Eurípedes de Marília (UNIVEM), Marília, Brasil. gustavo.camossi@unesp.br.

² Doutoranda em Ciência da Informação, Universidade Estadual Paulista (UNESP), Marília, São Paulo, Brasil. Bolsista da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP). af.jesus@unesp.br.

³ Doutoranda em Ciência da Informação pela Universidade Estadual Paulista (UNESP). suellen.timm@unesp.br.

⁴ Docente em Informação e Tecnologia pela Universidade de São Paulo (USP), Ribeirão Preto, São Paulo, Brasil. Professor Associado, Universidade de São Paulo (USP). santarem.segundo@unesp.br.

⁵ Doutora em Educação, Universidade Estadual Paulista Júlio de Mesquita Filho (Unesp), Marília, SP, Brasil. Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), Brasília, Brasil. silvana.vidotti@unesp.br.

baseada na análise de 10 artigos sobre o tema. Os resultados indicam que a Inteligência Artificial Generativa contribui para a eficiência e personalização em Experiência do Usuário, mas sua implementação requer planejamento estratégico e avaliação dos impactos nas necessidades dos usuários. Entre os principais desafios identificados, destacam-se a qualidade dos resultados, a ausência de diretrizes claras, questões relacionadas à privacidade e segurança, além da dependência da tecnologia, evidenciando a necessidade de regulamentação.

Palavras-chave: Inteligência Artificial Generativa; Experiência do Usuário; análise exploratória de literatura.

Abstract

Generative Artificial Intelligence (GAI) stands out in User Experience (UX) by automating processes, personalizing interfaces, and optimizing tests, but it also raises discussions about ethics and collaboration between humans and machines. This study aims to identify how Generative Artificial Intelligence (GAI) has been discussed and applied in User Experience (UX), highlighting its challenges and potentials. The research, with an exploratory and descriptive approach focused on qualitative analysis, examined 10 articles on the topic. The results indicate that GAI improves efficiency and personalization in UX but requires strategic planning and evaluation of its impacts on users' needs. Additionally, the main challenges include the quality of results, lack of clear guidelines, privacy and security issues, and technology dependence, highlighting the need for regulation.

Keywords: Generative Artificial Intelligence, User Experience, exploratory literature analysis.

1 INTRODUÇÃO

A Inteligência Artificial Generativa (IAG) é um ramo da Inteligência Artificial (IA) voltado para a criação de conteúdos contextualmente relevantes. Suas ferramentas podem gerar textos, imagens, vídeos, códigos e outros formatos de mídia, produzindo resultados que se assemelham significativamente aos criados por humanos ou coletados do mundo real (Banh; Strobel, 2023).

Essas ferramentas têm ganhado destaque como tecnologia inovadora, capaz de criar conteúdo complexo em diversos formatos, tendo seu funcionamento baseado no uso de grandes bases de dados. Sua aplicação no campo da Experiência do Usuário (UX) traz novos horizontes, não apenas na automação de processos de design, mas também na personalização de interfaces e na criação de soluções mais dinâmicas e adaptáveis (Johnston *et al.*, 2019; Oh *et al.*, 2018). No entanto, essa tecnologia também desperta debates sobre ética, transparência e desafios na colaboração entre humanos e máquinas (Koch, 2017; Oh *et al.*, 2018).

A UX é fundamental no desenvolvimento de soluções digitais, pois visa atender às necessidades e expectativas dos usuários, impactando diretamente a aceitação e o uso de sistemas. Pesquisas indicam que falhas na UX podem resultar na rejeição de tecnologias pelos usuários, frequentemente associadas a sentimento de frustração e desconfiança (Hart; Sutcliffe, 2019; Nisafani; Kiely; Mahony, 2020).

A incorporação da IAG no design UX apresenta potencial para minimizar tais problemas, otimizando processos e gerando soluções mais alinhadas às demandas dos usuários. Contudo, ainda existem lacunas significativas na literatura quanto à forma como essa integração ocorre e quais são suas limitações (Abbas; Ghauth; Ting, 2022; Wallach; Flohr; Kaltenhauser, 2020).

Estudos prévios indicam que a IAG tem sido aplicada em diversas áreas, incluindo a geração de protótipos, a personalização de interfaces e a avaliação automatizada de soluções (Feldman, 2017; Salminen *et al.*, 2019). No entanto, esses usos enfrentam desafios como a dificuldade de aplicabilidade dos sistemas de IA, a falta de controle por parte dos profissionais da área e a necessidade de habilidades especializadas para integrar tecnologias avançadas nos processos de design (Asatiani *et al.*, 2020; Pandian; Suleri, 2020).

Este estudo tem como objeto de pesquisa a aplicação da Inteligência Artificial Generativa no campo da Experiência do Usuário, tendo o objetivo de identificar como IAG tem sido discutida e aplicada na UX, destacando seus desafios e potenciais aplicações. A problemática central que orienta esta investigação é: Como a Inteligência Artificial Generativa tem sido discutida e aplicada no contexto da Experiência do Usuário? A partir dessa questão, o estudo propõe mapear as discussões existentes, identificar as principais aplicações práticas e evidenciar lacunas no conhecimento que possam direcionar pesquisas futuras.

A revisão inicial a respeito da temática revelou uma escassez de estudos focados no assunto, o que justifica a adoção de uma abordagem exploratória em vez de uma abordagem sistemática. A Análise Exploratória proporciona maior flexibilidade na seleção das fontes e dos estudos que irão compor o *corpus* da pesquisa, sendo frequentemente utilizada como etapa inicial para estudos sistemáticos, necessária especialmente em casos de escassez de estudos consolidados na literatura científica a respeito do tema. Além disso, a pesquisa foi orientada por um protocolo de pesquisa e pelo uso de ferramentas tecnológicas, visando maior precisão e eficiência na coleta de dados.

2 PROCEDIMENTOS METODOLÓGICOS

Do ponto de vista metodológico, esta pesquisa é de natureza exploratória, pois busca reunir informações preliminares, aprofundar a compreensão teórica e estabelecer uma base

para investigações futuras (Prodanov; Freitas, 2013). Além disso, possui um caráter descritivo, especialmente na análise da ferramenta de apoio utilizada. Quanto à abordagem dos resultados, trata-se de um estudo qualitativo.

Para alcançar o objetivo proposto, foi realizada uma análise exploratória da literatura científica, combinada com a técnica de *snowballing*, que consiste em avaliar as listas de referências e de citações de estudos, a fim de identificar novos estudos relevantes para a pesquisa (Silva, 2017, p. 29). Toda a pesquisa foi conduzida por um protocolo de pesquisa pré-estabelecido, conforme apresentado no Quadro 1.

Quadro 1 – Protocolo de pesquisa

Protocolo de pesquisa	
Pergunta de pesquisa (principal)	Como a Inteligência Artificial Generativa tem sido discutida e aplicada no contexto da Experiência do Usuário?
Objetivos	Identificar como Inteligência Artificial Generativa tem sido discutida e aplicada na Experiência do Usuário, destacando seus desafios e potenciais aplicações
Palavras-chave	("generative artificial intelligence" OR GenAI) AND ("user experience" OR UX)
Bases de dados consultada	WOS; Google Acadêmico e semantic scholar (via elicit)
Estratégia de busca	Uso dos operadores booleanos AND e OR para elaboração da estratégia de busca e aplicação de filtros para: artigos e artigos de revisão na <i>Web of Science</i>
Período abrangido	sem recorte temporal
Idiomas	inglês
Crítérios de Inclusão	(I) Discute a relação entre Inteligência Artificial Generativa e Experiência do Usuário; (I) Aplica a Inteligência Artificial Generativa no âmbito da Experiência do Usuário
Crítérios de exclusão	(E) Não está no idioma estabelecido para a pesquisa; (E) Apenas menciona a temática de interesse; (E) Não aborda a temática de interesse; (E) Não foi possível ter acesso ao documento completo
Formulário de extração	1) Objetivo do artigo; 2) Potenciais usos; 3) contexto da aplicação realizada; 4) desafios e problemáticas mencionados

Fonte: Elaboração própria.

Apresentados os procedimentos metodológicos adotados, a próxima seção detalha os resultados obtidos.

3 RESULTADOS

Esta seção apresenta os resultados obtidos a partir da aplicação das estratégias de busca e seleção descritas no protocolo de pesquisa. No total, foram recuperados 47

documentos, dos quais 35 foram descartados, 1 foi identificado como duplicado e 10 foram aceitos para compor o *corpus* da pesquisa.

O Quadro 2 apresenta a lista de documentos aceitos, acompanhada de uma breve descrição de seus objetivos.

Quadro 2 - Documentos aceitos para compor o *corpus* da pesquisa

Autores	Título	Objetivos
Houde <i>et al.</i> (2022)	Opportunities for Generative AI in UX Modernization 81-91	Explora e identifica oportunidades para o uso de Inteligência Artificial Generativa na modernização da Experiência do Usuário, particularmente em aplicações legadas.
Chan (2023)	The Consistency between Popular Generative Artificial Intelligence (AI) Robots in Evaluating the User Experience of Mobile Device Operating Systems	Analisa as avaliações realizadas por dois robôs de inteligência artificial generativa, Dragonfly e GPT-4, considerando a experiência do usuário em iOS e Android.
Ayemowa, Ibrahim e Bena (2024)	A Systematic Review of the Impact of Auxiliary Information on Recommender Systems	Realiza uma revisão sistemática sobre o impacto da incorporação de informações auxiliares nos sistemas de recomendação, analisa os avanços recentes, metodologias utilizadas, bases de dados, métricas de avaliação e explora a significância dessas informações no contexto da Inteligência Artificial Generativa.
Chen (2024)	Optimization of User Experience Design Process Driven by AIGC	Explorar como a Inteligência Artificial Generativa pode otimizar o processo de design da Experiência do Usuário (UX), tornando-o mais eficiente e adaptável às necessidades dos usuários.
Huang, Kanij e Mahajan (2024)	Unlocking Adaptive User Experience with Generative AI	Investiga o uso do ChatGPT para criar personas e interfaces adaptativas, comparando com métodos manuais.
Kalva (2024)	Optimizing E-commerce Platforms with GenAI-Driven DevOps and LLMOps A Scalable Framework for Enhanced User Experience	Discute como a personalização impulsionada por IA pode melhorar as recomendações personalizadas em plataformas de e-commerce, enfrentando desafios técnicos e éticos.
Li <i>et al.</i> (2024)	User Experience Design Professionals' Perceptions of Generative Artificial Intelligence	Busca, a partir de entrevistas, avaliar as expectativas e apreensões de User Experience Designers a respeito do uso de IAG.
Sikorski (2024)	Leveraging Generative AI Tools for UX Design in Lean and Agile Projects	Propõe um framework de decisão para gerentes de projetos de TI e profissionais de UX sobre o uso de ferramentas de IA generativa em projetos ágeis e <i>lean</i> .
Takaffoli, Li e Mäkelä (2024)	Generative AI in User Experience Design and Research: How Do UX Practitioners, Teams, and Companies Use GenAI in Industry?	Investiga como a Inteligência Artificial Generativa está sendo utilizada na prática de UX (User Experience) na indústria. O estudo busca preencher lacunas na pesquisa ao explorar práticas individuais, de equipe e organizacionais relacionadas ao uso da IA Generativa em UX.
Weisz <i>et al.</i> (2024)	Design Principles for Generative AI Applications	Desenvolve e refina princípios de design para aplicações que incorporam Inteligência Artificial Generativa, com o intuito de criar diretrizes claras e

		eficazes para o desenvolvimento de sistemas baseados nessa tecnologia.
--	--	--

Fonte: Elaboração própria.

A análise do Quadro 2 revela que, apesar das diferenças entre os documentos, todos têm em comum a busca por discutir as oportunidades e as problemáticas do uso de IAG em UX. As abordagens exploram o impacto dessa tecnologia na indústria, nos usuários e nos profissionais de design de UX. Destacam-se, ainda, estudos com abordagens aplicadas, que buscam propor instrumentos e avaliar ferramentas para tarefas específicas de UX, como os trabalhos de Sikorski (2024); Huang, Kanij e Mahajan (2024) e Chan (2023).

Identificados os estudos e seus objetivos, a próxima etapa da pesquisa foi a coleta das potenciais aplicações de IAG em UX, bem como os desafios e as problemáticas relacionadas a essa aplicação. As próximas subseções apresentam os resultados dessa análise.

3.1 Oportunidades Relacionadas ao uso de IAG em UX

Para compreender como a literatura aborda as oportunidades associadas ao uso de Inteligência Artificial Generativa em Experiência do Usuário, foram identificados e sistematizados os benefícios mencionados pelos autores. A organização dessas oportunidades no Quadro 3 possibilita a identificação dos aspectos mais recorrentes na literatura, destacando tendências e aplicações potenciais.

Quadro 3 - Oportunidades relacionadas ao uso de IAG em UX

Oportunidade	Autores
Automação de tarefas repetitivas no design	Li <i>et al.</i> (2024); Ayemowa, Ibrahim e Bena (2024); Takaffoli, Li e Mäkelä (2024); Houde <i>et al.</i> (2022); Chan (2023)
Geração de interfaces e conteúdos personalizados	Li <i>et al.</i> (2024); Houde <i>et al.</i> (2022); Chen (2024)
Otimização de testes de usabilidade	Li <i>et al.</i> (2024); Takaffoli, Li e Mäkelä (2024); Chen (2024)
Apoio à tomada de decisão em UX	Li <i>et al.</i> (2024); Ayemowa, Ibrahim e Bena (2024); Houde <i>et al.</i> (2022); Chan (2023)
Melhoria na acessibilidade e inclusão digital	Li <i>et al.</i> (2024); Ayemowa, Ibrahim e Bena (2024); Chen (2024)
Redução do tempo no desenvolvimento de interfaces	Li <i>et al.</i> (2024); Houde <i>et al.</i> (2022)
Possibilidade de explorar múltiplas alternativas de design	Li <i>et al.</i> (2024); Chen (2024); Takaffoli, Li e Mäkelä (2024)

Integração com sistemas de recomendação para personalização dinâmica	Li et al. (2024); Ayemowa, Ibrahim e Bena (2024)
Redução de barreiras técnicas no design para profissionais não especialistas	Li et al. (2024); Chan (2023)
Aprimoramento da colaboração entre equipes de UX e desenvolvimento	Li et al. (2024); Houde et al. (2022); Chen (2024)

Fonte: Elaboração própria.

A análise do Quadro 3 indica que a automação de tarefas repetitivas no design é uma das aplicações mais recorrentes da IAG. Esse uso pode ser observado em atividades como geração de *wireframes*, prototipagem de baixa fidelidade e testes de acessibilidade. Estudos apontam que a automação dessas etapas permite otimizar fluxos de trabalho e reorganizar as atividades desempenhadas por profissionais de UX (Li et al., 2024; Ayemowa; Ibrahim; Bena, 2024; Takaffoli; Li; Mäkelä, 2024; Houde et al., 2022; Chan, 2023).

Outro aspecto relevante é a personalização dinâmica da experiência do usuário, que consiste na adaptação de interfaces e conteúdos com base nas interações do usuário. Modelos de IA aplicados a sistemas de recomendação possibilitam ajustes em tempo real, propiciando maior engajamento e uma experiência mais alinhada às preferências individuais (Li et al., 2024; Houde et al., 2022; Chen, 2024).

A otimização de testes de usabilidade é outra aplicação frequentemente abordada na literatura. Ferramentas de IAG podem ser utilizadas para analisar interações e detectar padrões de uso, auxiliando na identificação de ajustes em interfaces digitais (Li et al., 2024; Takaffoli; Li; Mäkelä, 2024; Chen, 2024). Esse processo pode facilitar a realização de iterações durante o desenvolvimento de produtos digitais.

A tomada de decisão em UX também é um aspecto mencionado, indicando que a IAG pode ser utilizada para gerar análises preditivas e apresentar recomendações baseadas em padrões de comportamento dos usuários (Li et al., 2024; Ayemowa; Ibrahim; Bena, 2024; Houde et al., 2022; Chan, 2023). Essa aplicação pode ser integrada a processos estratégicos, fornecendo insumos para ajustes e refinamentos de interfaces.

A literatura destaca ainda a melhoria na acessibilidade e inclusão digital, considerando que modelos de IA podem auxiliar na adaptação automática de interfaces para diferentes perfis de usuários. Isso inclui ajustes de contraste, descrição de imagens e recomendações para aprimorar a usabilidade em conformidade com normas de acessibilidade (Li et al., 2024; Ayemowa; Ibrahim; Bena, 2024; Chen, 2024).

A redução do tempo no desenvolvimento de interfaces é outro ponto mencionado, indicando que a IAG pode ser empregada para converter esboços em código funcional e otimizar a colaboração entre designers e desenvolvedores (Li *et al.*, 2024; Houde *et al.*, 2022).

Além disso, a literatura sugere que a exploração de múltiplas alternativas de design pode ser facilitada por IAG, permitindo a geração automática de diferentes variações de interfaces para análise e seleção pelos profissionais envolvidos no processo de UX (Li *et al.*, 2024; Chen, 2024; Takaffoli; Li; Mäkelä, 2024).

A integração da IAG com sistemas de recomendação também é apontada como uma possibilidade de uso, uma vez que a IA pode aprimorar a personalização da experiência do usuário ao analisar grandes volumes de dados e ajustar conteúdos e funcionalidades conforme padrões de interação (Li *et al.*, 2024; Ayemowa; Ibrahim; Bena, 2024).

Outro aspecto identificado é a redução de barreiras técnicas no design para profissionais sem formação especializada, permitindo que ferramentas de IA auxiliem no desenvolvimento de interfaces digitais sem a necessidade de conhecimentos avançados em programação ou design (Li *et al.*, 2024; Chan, 2023).

Por fim, destaca-se que IAG pode aprimorar a colaboração entre equipes de UX e desenvolvimento, uma vez que essas ferramentas podem facilitar a comunicação entre as diferentes áreas envolvidas no processo de criação de interfaces (Li *et al.*, 2024; Houde *et al.*, 2022; Chen, 2024).

Em síntese, os estudos analisados indicam que a IAG pode contribuir para a automação de processos, a personalização de interfaces e a eficiência no desenvolvimento de produtos digitais. Contudo, a implementação dessas aplicações requer um planejamento estratégico para garantir que a adoção da tecnologia esteja alinhada às necessidades e objetivos do projeto de UX.

3.2 Desafios e Problemáticas Relacionados ao uso de IAG em UX

Buscando entender como os autores estão discutindo o uso de IAG em UX, foram coletados nos documentos os desafios relacionados a essa aplicação. O Quadro 4 apresenta os resultados dessa análise.

Quadro 4 - Análise dos desafios e problemáticas relacionados ao uso de IAG em UX

Desafio	Autores
Preocupação com a qualidade dos resultados	Li <i>et al.</i> (2024); Ayemowa, Ibrahim e Bena (2024); Takaffoli, Li e Mäkelä (2024); Houde <i>et al.</i> (2022); Chen (2024); Chan (2023)
Necessidade de validação humana	Li <i>et al.</i> (2024); takaffoli, Li e Mäkelä (2024); Chen (2024)
Preocupação com a violação de direitos autorais e propriedade	Li <i>et al.</i> (2024); Chen (2024)
Degradação de habilidades dos profissionais, afetando a sua formação	Li <i>et al.</i> (2024); Houde <i>et al.</i> (2022); Chen (2024)
Preocupação com a Privacidade e segurança dos dados	Li <i>et al.</i> (2024); Ayemowa, Ibrahim e Bena (2024); Houde <i>et al.</i> (2022); Chen (2024)
Uso de dados enviesados no treinamento	Li <i>et al.</i> (2024); Ayemowa, Ibrahim e Bena (2024); Houde <i>et al.</i> (2022); Chan (2023); Huang, Kanij e Mahajan (2024)
Dificuldade na avaliação dos resultados	Li <i>et al.</i> (2024); Huang, Kanij e Mahajan (2024)
Dificuldade para atribuição da autoria dos resultados	Li <i>et al.</i> (2024)
Custos computacionais elevados	Ayemowa, Ibrahim e Bena (2024)
A personalização excessiva pode limitar a diversidade de conteúdos	Ayemowa, Ibrahim e Bena (2024)
Inconsistências entre diferentes modelos de IA	Chan (2023)
Preocupação com a falta de transparência	Li <i>et al.</i> (2024); Houde <i>et al.</i> (2022); Chen (2024)
Falta de Políticas e Regulamentação para o Uso da IA	Houde <i>et al.</i> (2022); Chen (2024)
Desafios no Design de Interfaces Complexas	Huang, Kanij e Mahajan (2024)
Dependência de Engenharia de Prompt	Huang, Kanij e Mahajan (2024)

Fonte: Elaboração própria.

Observa-se um destaque na preocupação com a qualidade dos resultados obtidos, sendo elencados problemas como resultados genéricos, imprecisos, incorretos, pouco confiáveis; incompreensíveis para o usuário ou ainda pouco alinhados com as necessidades/objetivos dos usuários (Li *et al.*, 2024; Ayemowa; Ibrahim; Bena, 2024; Takaffoli; Li; Mäkelä, 2024; Houde *et al.*, 2022; Chen, 2024; Chan, 2023).

Esse desafio se relaciona ainda com as dificuldades na avaliação dos resultados obtidos. Li *et al.*, (2024) mencionam a falta de diretrizes para orientar esse processo de avaliação no contexto da UX. Huang, Kanij e Mahajan (2024) apontam que a avaliação qualitativa, muitas vezes adotada nesse contexto, possui um caráter subjetivo. Os autores mencionam, ainda, a dificuldade de generalização dos resultados de avaliação dessas ferramentas, já que a avaliação de qualidade possui uma influência fortemente contextual. Justamente por essa problemática em torno da qualidade dos resultados, as ferramentas exigem a constante validação humana, fazendo com que nem sempre o uso dessas ferramentas signifique economia de tempo.

Ainda relacionado à qualidade dos resultados, tem-se a preocupação com o uso de dados enviesados, pouco diversos ou culturalmente consistentes (Li *et al.*, 2024; Ayemowa; Ibrahim; Bena, 2024; Houde *et al.*, 2022; Chan, 2023; Huang; Kanij; Mahajan, 2024). Como esses dados são utilizados no treinamento das IAG e tem impacto direto nos resultados obtidos, destaca-se a preocupação com resultados com caráter discriminatório.

Um outro aspecto abordado pelos autores é o risco de degradação das habilidades dos profissionais de UX, decorrentes de uma possível dependência excessiva dessas ferramentas. Essa precaução se ressalta quando relacionada a novos profissionais, que, por essa dependência, podem não desenvolver as competências e habilidades necessárias para o crescimento e evolução na carreira (Li *et al.*, 2024; Houde *et al.*, 2022; Chen, 2024).

Entre os desafios elencados, sobressaem as preocupações com a privacidade dos dados, a segurança das informações compartilhadas e o risco de violação inconsciente de direitos autorais (Li *et al.*, 2024; Ayemowa; Ibrahim; Bena, 2024; Houde *et al.*, 2022; Chen, 2024). Esse problema se agrava devido às ausências de regulamentação específica para o uso da IA e de políticas internas para orientar os profissionais a respeito do uso seguro e ético de IAG (Houde *et al.*, 2022; Chen, 2024).

Além disso, a transparência limitada sobre o treinamento do funcionamento das IAG, dificulta o entendimento sobre como essas ferramentas estão lidando com aspectos de privacidade e direitos autorais, tornando o desafio ainda mais crítico (Li *et al.*, 2024; Houde *et al.*, 2022; Chen, 2024).

A IAG tem se inserido como ferramentas de suporte em diversos contexto de atuação. Como discutido na seção 3.1, elas têm o potencial de contribuir para a automação de

atividades repetitivas do processo de design, de permitir que diversas soluções sejam exploradas para um mesmo problema, de reduzir barreiras técnicas para novos profissionais e de auxiliar na criação de experiências personalizadas que levem em consideração aspectos de acessibilidade. Entretanto, a análise dos estudos permite compreender que, no estágio atual das ferramentas de IAG, uma série de desafios impedem que todo esse potencial seja explorado.

Um dos principais aspectos que dificultam a adoção de IAG no contexto de UX é a dificuldade de alcançar os níveis de qualidade exigidos para que essas ferramentas possam efetivamente contribuir para as práticas da área. Muitas vezes, os resultados obtidos ainda são genéricos, não se aprofundam nas necessidades dos usuários e apresentam o desafio adicional de lidar com inconsistências, mesmo quando se utiliza o mesmo *prompt*.

Embora possa contribuir para a atuação de novos profissionais, reduzindo as barreiras de conhecimento técnicos, o uso de IAG depende justamente do conhecimento dos profissionais para validar a relevância dos resultados obtidos, o que demonstra ser um grande desafio, ressaltado pela ausência de diretrizes e orientações claras sobre como avaliar esses resultados. O desafio do uso de UX por novos profissionais levanta ainda a preocupação sobre a degradação das habilidades dos novos profissionais, levantando a necessidade de pesquisas sobre como garantir que a formação plena desse profissional não seja afetada pelo uso indiscriminado de IAG.

Com base na análise dos desafios apresentados no Quadro 4, foi possível compreender que muitos desses estão relacionados a problemas informacionais. Esses desafios envolvem a falta de transparência quanto ao funcionamento das ferramentas, questões relacionadas à curadoria responsável e ética dos dados utilizados em seu treinamento, às políticas de privacidade, às estratégias de proteção de dados sensíveis e ao respeito aos direitos autorais.

A IAG está transformando a área de UX ao introduzir novas possibilidades e desafios. Essa transformação não se limita apenas às ferramentas e processos de design, mas também exige uma reflexão sobre como essa tecnologia impactará a forma como os usuários interagem com a informação e os sistemas digitais. Em termos de tendências interdisciplinares, a integração da IAG em UX cria um espaço de convergência entre Ciência da Informação, Ciência da Computação, Design e Psicologia.

A Ciência da Informação, em particular, desempenha um papel na organização, representação, acesso e uso da informação gerada e utilizada por sistemas de IAG, tendo o potencial de contribuir para a discussão dos diversos problemas informacionais. A capacidade da IAG de gerar grandes volumes de conteúdo personalizado levanta questões importantes para a Ciência da Informação, como a necessidade de desenvolver novos métodos de curadoria, avaliação da qualidade e confiabilidade da informação, e estratégias para garantir o acesso equitativo e a preservação da informação gerada por essas ferramentas. Além disso, a Ciência da Informação pode contribuir para a criação de interfaces mais intuitivas e acessíveis, que facilitem a interação dos usuários com sistemas de IAG, promovendo uma Experiência do Usuário mais eficaz e satisfatória.

Assim, o impacto da IAG na Ciência da Informação se estende para além da organização e recuperação da informação. A automatização de tarefas de design e a geração de conteúdo por sistemas de IAG podem transformar os fluxos de trabalho dos profissionais da informação, exigindo novas habilidades e competências. A Ciência da Informação precisa se adaptar a essas mudanças, formando profissionais capazes de trabalhar em conjunto com sistemas de IAG, desenvolvendo métodos para avaliar criticamente a informação gerada por essas ferramentas e criando soluções inovadoras para os desafios informacionais da era da Inteligência Artificial Generativa.

4 CONSIDERAÇÕES FINAIS

A Inteligência Artificial Generativa tem se mostrado uma ferramenta relevante na Experiência do Usuário (UX), contribuindo para a automação de processos, como a criação de *wireframes*, prototipagens e testes de acessibilidade, além de personalização de interfaces e otimização de testes de usabilidade. Apesar dos benefícios, a adoção dessa tecnologia suscita discussões sobre ética, confiabilidade e colaboração entre humanos e máquinas no design de UX. Neste contexto, este estudo visa explorar essas questões, mapear aplicações práticas da IAG em UX e identificar lacunas para futuras investigações.

A análise de um conjunto de 10 documentos revelou que, embora a automação e a personalização promovam maior eficiência e inovação no design de UX, os estudos ressaltam a necessidade de um planejamento estratégico para alinhar a adoção da IAG aos objetivos do

projeto. Além disso, a colaboração entre equipes e a melhoria da acessibilidade são benefícios destacados. No entanto, torna-se necessário avaliar os impactos dessas tecnologias nas necessidades e expectativas dos usuários, a fim de mitigar riscos e maximizar a efetividade.

Entre os principais desafios, destacam-se a variabilidade na qualidade e confiabilidade dos resultados gerados pela IAG, a ausência de diretrizes normativas para avaliação de suas aplicações e questões relacionadas à privacidade e segurança dos dados. Adicionalmente, a dependência excessiva dessa tecnologia por profissionais de UX e a ausência de transparência nos modelos de IA reforçam a necessidade de regulamentação para garantir o uso ético e seguro. Estudos futuros podem aprofundar a análise dos impactos técnicos e éticos da IAG na UX, além da formulação de políticas para o uso responsável e eficiente.

REFERÊNCIAS

- ABBAS, A. M. H.; GHAUTH, K. I.; TING, C. User experience design using machine learning: a systematic review. **Leee Access**, [S. l.], v. 10, p. 51501-51514, 2022. Disponível em: <http://dx.doi.org/10.1109/access.2022.3173289>. Acesso em: 23 fev. 2025.
- ASATIANI, A. *et al.* Challenges of explaining the behavior of black-box AI systems. **MIS Quarterly Executive**, v. 19, n. 4, p. 259-278, 2020. Disponível em: <https://ieeexplore.ieee.org/document/10681332>. Acesso em: 23 fev. 2025.
- AYEMOWA, M.; IBRAHIM, R.; BENA, Y. A. A systematic review of the impact of auxiliary information on recommender systems. **IEEE Access**, [S. l.], v. 12, p. 139524-139539, 2024. Disponível em: <http://dx.doi.org/10.1109/access.2024.3462750>. Acesso em: 23 fev. 2025.
- BANH, L.; STROBEL, G. Generative Artificial Intelligence. **Electronic Markets**, v. 33, n. 1, p. 63, 2023. Disponível em: <https://link.springer.com/article/10.1007/s12525-023-00680-1>. Acesso em: 07 jul. 2024.
- CHAN, V. The consistency between popular generative Artificial Intelligence (AI) robots in evaluating the user experience of mobile device operating systems. **Ahfe International**, [S. l.], v. 113, n. 1, p. 205-215, 2023. Disponível em: <http://dx.doi.org/10.54941/ahfe1004193>. Acesso em: 23 fev. 2025.
- CHEN, S. Optimization of user experience design process driven by AIGC. **Frontiers in computing and intelligent systems**, [S. l.], v. 9, n. 1, p. 20-22, 29 jul. 2024. Disponível em: <http://dx.doi.org/10.54097/wwermb12>. Acesso em: 23 fev. 2025.
- FELDMAN, S. Co-Creation: human and ai collaboration in creative expression. **Electronic Workshops in Computing**, [S. l.], p. 422-429, 2017. Disponível em: <http://dx.doi.org/10.14236/ewic/eva2017.84>. Acesso em: 23 fev. 2025.

HART, J.; SUTCLIFFE, A. Is it all about the apps or the device?: user experience and technology acceptance among ipad users. **International Journal of Human-Computer Studies**, [S. l.], v. 130, p. 93-112, out. 2019. Disponível em: <http://dx.doi.org/10.1016/j.ijhcs.2019.05.002>. Acesso em: 23 fev. 2025.

JOHNSTON, Vivien *et al.* A framework for the development of a dynamic adaptive intelligent user interface to enhance the user experience. *In*: EUROPEAN CONFERENCE ON COGNITIVE ERGONOMICS, 31., 2019, United Kingdom. **Proceedings [...]**. New York: Association for Computing Machinery, 2019. p. 32-35. Disponível em: <http://dx.doi.org/10.1145/3335082.3335125>. Acesso em: 2 fev. 2025.

KALVA, R. Optimizing e-commerce platforms with GenAI-Driven DevOps and LLMOps a scalable framework for enhanced user experience. **Journal of Artificial Intelligence, Machine Learning and Data Science**, [S. l.], v. 2, n. 4, p. 1782-1788, dez. 2024. Disponível em: <http://dx.doi.org/10.51219/jaimld/rahul-kalva/396>. Acesso em: 23 fev. 2025.

KOCH, Janin. Design implications for designing with a collaborative AI. *In*: AAAI SPRING SYMPOSIA, 2017, [S. l.]. **Anais [...]**. [S. l.]: AAAI, 2017. Disponível em: <https://cdn.aaai.org/ocs/15382/15382-68259-1-PB.pdf>. Acesso em: 02 fev. 2025.

LI, Jie *et al.* User Experience design professionals' perceptions of generative Artificial Intelligence. *In*: CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS, 24., 2024, Honolulu. **Proceedings [...]**. New York: Association for Computing Machinery, 2024. p. 1-18. Disponível em: <http://dx.doi.org/10.1145/3613904.3642114>. Acesso em: 23 fev. 2024.

NISAFANI, A. S.; KIELY, G.; MAHONY, C. Workers' technostress: a review of its causes, strains, inhibitors, and impacts. **Journal Of Decision Systems**, [S. l.], v. 29, n. 1, p. 243-258, 26 jul. 2020. Informa UK Limited. Disponível em: <http://dx.doi.org/10.1080/12460125.2020.1796286>. Acesso em: 07 jul. 2024.

OH, C. *et al.* I Lead, you help but only with enough details. *In*: CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS, 18., 2018, Montreal. **Proceedings [...]**. New York: Association for Computing Machinery, 2018. p. 1-13. Disponível em: <http://dx.doi.org/10.1145/3173574.3174223>. Acesso em: 02 fev. 2025.

PANDIAN, V. P. S.; SULERI, S. BlackBox toolkit: intelligent assistance to UI design. **arXiv preprint arXiv:2004.01949**. 2020. Disponível em: <https://auld.aaai.org/Press/Reports/Symposia/Spring/ss-17.php>. Acesso em: 20 maio 2025.

PRODANOV, C.; FREITAS, E. **Metodologia do trabalho científico**: métodos e técnicas da pesquisa e do trabalho acadêmico. 2. ed. Novo Hamburgo: Editora Feevale, 2013. 277 p.

SALMINEN, J.; SENGÜN, S.; JUNG, S.; JANSEN, B. J. Design issues in automatically generated persona profiles. *In*: CONFERENCE ON HUMAN INFORMATION INTERACTION AND RETRIEVAL, 19., 2019, Glasgow. **Proceedings [...]**. New York: Association for Computing Machinery, 2019. p. 225-229. Disponível em: <http://dx.doi.org/10.1145/3295750.3298942>. Acesso em: 07 jul. 2024.

SIKORSKI, M. Leveraging generative AI tools for UX design in Lean and Agile Projects. *In: INTERNATIONAL CONFERENCE ON INFORMATION SYSTEMS DEVELOPMENT*, 32., 2024, Poland. **Proceedings [...]**. Poland: [S. l.], 2024. p. 1-4. Disponível em: <https://aisel.aisnet.org/isd2014/proceedings2024/agile/4/>. Acesso em: 23 fev. 2025.

TAKAFFOLI, M.; LI, S.; MÄKELÄ, V. Generative AI in user experience design and research: how do ux practitioners, teams, and companies use genai in industry? *In: DESIGNING INTERACTIVE SYSTEMS CONFERENCE*, 24., 2024, Copenhagen. **Proceedings [...]**. New York: Association for Computing Machinery, 2024. p. 157. Disponível em: <https://doi.org/10.1145/3643834.3660720>. Acesso em: 23 fev. 2025.

WALLACH, D. P.; FLOHR, L. A.; KALTENHAUSER, A. Beyond the buzzwords: on the perspective of ai in ux and vice versa. **Lecture Notes li Computer Science**, [S. l.], p. 146-166, 2020. Springer International Publishing. Disponível em: http://dx.doi.org/10.1007/978-3-030-50334-5_10. Acesso em: 23 fev. 2025.

WEISZ, J. D. *et al.* Design principles for generative AI applications. *In: CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS*, 24., 2014, Honolulu. **Proceedings [...]**. New York: Association for Computing Machinery, 2024. p. 1-22. Disponível em: <http://dx.doi.org/10.1145/3613904.3642466>. Acesso em: 23 fev. 2025.



[4.0 Internacional](https://creativecommons.org/licenses/by/4.0/)

Esta licença permite compartilhamento, remixe, adaptação e criação a partir do trabalho, mesmo para fins comerciais, desde que sejam atribuídos créditos ao(s) autor(es). Conteúdos elaborados por terceiros, citados e referenciados nesta obra não são cobertos pela licença.

Como citar este trabalho:

CAMOSSÍ, Gustavo; JESUS, Ananda Fernanda de; BARROS, Suellen Timm; SANTAREM SEGUNDO, José Eduardo; VIDOTTI, Silvana Aparecida Borsetti Gregorio. Aplicação da Inteligência Artificial Generativa na Experiência do Usuário: potencialidades e desafios. *In: WORKSHOP DE INFORMAÇÃO DADOS E TECNOLOGIA*, 8., 2025, Marília, SP. **Anais [...]**. Marília, SP: Universidade de Marília, 2025. DOI: <http://doi.org/10.22477/viii.widat.233>.



WIDaT 2025

VIII WORKSHOP DE INFORMAÇÃO,
DADOS E TECNOLOGIA

MARÍLIA • SP

Bases de dados qualificadas para a descoberta de conhecimentos úteis à gestão de acervos culturais

Qualified databases for discovering knowledge to support the management of cultural collections

Daniela Lucas da Silva Lemos¹



<http://lattes.cnpq.br/9280443047358807>



<https://orcid.org/0000-0003-1565-7366>

Henrique Monteiro Cristovão²



<http://lattes.cnpq.br/5035919384923489>



<https://orcid.org/0000-0003-2011-7022>

Resumo

Padrões de documentação são usados amplamente nas ações de organização e tratamento da informação para a constituição de bases de dados com qualidade, na intenção de se obter informações estratégicas para a gestão de acervos culturais. O presente estudo relata a experiência com a construção das bases de dados do acervo do Fundo de Cultura do Espírito Santo, sob a gestão da Secretaria de Estado da Cultura, disponibilizadas na rede de provedores da plataforma MEDIATECA Capixaba. Assim, a partir da análise exploratória destas bases, a pesquisa objetivou identificar agrupamentos coesos nos autores contemplados nos editais por meio de similaridades nos eixos temáticos de seus respectivos projetos. Metodologicamente, a modelagem de metadados, incluindo o uso de vocabulários controlados e do padrão de metadados *Dublin Core*, contou com a técnica conhecida como Matriz de Restrições para a elaboração de perfis de aplicação para a constituição das bases de dados qualificadas. Por conseguinte, as redes de informações projetadas a partir da exploração de dados dessas bases foram construídas a partir do método de Análise de Redes Sociais. A análise de redes de informação revelou relações entre autores, editais e temas, evidenciando padrões de participação e distribuição de recursos. Além disso, a análise exploratória de dados permitiu identificar tendências para subsidiar políticas públicas culturais. Assim, a pesquisa reforça a importância dos padrões de documentação para qualificar os dados e contribuir para a gestão da informação cultural em ambientes digitais, indicando caminhos para aprimoramentos na organização de acervos e na democratização do acesso à cultura.

Palavras-chave: padrões de documentação; acervos culturais; qualidade de dados; análise de redes de informação; descoberta de conhecimento.

Abstract

Documentation standards are widely used in information organization and processing actions to create quality databases, with the intention of obtaining strategic information for the management of cultural collections. This study reports the experience with the construction of databases of the Espírito Santo Culture Fund collection, under the management of the State Secretariat of Culture, made available on the

¹ Doutora em Ciência da Informação, Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, MG, Brasil. Professora e Pesquisadora, Universidade Federal do Espírito Santo (UFES), Vitória, ES, Brasil. daniela.l.silva@ufes.br.

² Doutor em Ciência da Informação, Universidade de Brasília (UnB), Brasília, DF, Brasil. Professor, Universidade Federal do Espírito Santo (UFES), Vitória, ES, Brasil. henrique.cristovao@ufes.br.

network of providers of the MEDIATECA Capixaba platform. Thus, based on the exploratory analysis of these databases, the research aimed to identify cohesive groupings in the authors contemplated in the calls for proposals through similarities in the thematic axes of their respective projects. Methodologically, the metadata modeling, including the use of controlled vocabularies and the Dublin Core metadata standard, relied on the technique known as the Restriction Matrix to develop application profiles for the creation of qualified databases. Consequently, the information networks designed from the exploration of data from these databases were built using the Social Network Analysis method. The analysis of information networks revealed relationships between authors, notices and themes, highlighting patterns of participation and distribution of resources. In addition, the exploratory data analysis allowed the identification of trends to support cultural public policies. Thus, the research reinforces the importance of documentation standards to qualify data and contributes to the management of cultural information in digital environments, indicating paths for improvements in the organization of collections and the democratization of access to culture.

Keywords: documentation standards; cultural collections; data quality; information network analysis; knowledge discovery.

1 INTRODUÇÃO

A padronização é um elemento crucial, discutido amplamente na Ciência da Informação, especialmente na área de Organização e Representação da Informação e do Conhecimento. Padrões de documentação são usados nessa perspectiva para apoiar práticas de catalogação e são essenciais na crescente digitalização de coleções de objetos culturais (Gilliland, 2016). Na plataforma digital de agregação de dados culturais, denominada MEDIATECA Capixaba (Lemos; Martins; Macedo, 2024), os padrões de documentação são fundamentados em teorias de catalogação descritiva e de assunto, focando na criação e no uso de metadados, vocabulários controlados e regras de catalogação pertinentes a cada perfil de aplicação.

O presente estudo relata a experiência com as bases de dados do acervo do Fundo de Cultura do Espírito Santo (Funcultura), sob a gestão da Secretaria de Estado da Cultura (SECULT/ES), implementadas em um dos repositórios institucionais da rede de provedores da MEDIATECA Capixaba, para a qual se buscou qualificar seus dados a partir de padrões de documentação na intenção de se obter informações e conhecimentos úteis à gestão de seu acervo cultural.

O Funcultura foi criado pela Lei Complementar Estadual nº458 de 21/10/2008 e, desde então, é o principal mecanismo do Governo do Estado para o fomento e incentivo a programas, projetos e ações culturais (Espírito Santo, 2023). Os recursos do Fundo são aplicados em projetos, ações e programas baseados em três dimensões culturais: cultura como expressão simbólica de um povo, cultura como direito e cidadania, e cultura como

economia que gera renda e trabalho. Nesse ensejo, há uma necessidade real de otimização do acesso à cultura frente ao mundo contemporâneo, com sua dinâmica digital que rege o cotidiano da sociedade e de grande parte dos cidadãos. Assim, a entrada deste acervo no projeto Midiateca Capixaba é justamente para oportunizar estudo, análise, pesquisa e organização do conjunto de dados acerca dos editais, projetos e produtos culturais em bases de dados qualificadas, possibilitando a sua exploração para fins de difundir informações úteis à sociedade bem como conhecimentos que permitam aos gestores tomarem decisões pertinentes para a elaboração de políticas públicas que sejam de interesse ao cidadão capixaba, em especial.

A análise de redes de informação tem sido usada para a projeção visual de descoberta de conhecimento em bases de dados e, principalmente, por ser uma ferramenta metodológica que viabiliza enxergar o que outras abordagens não permitem (Wasserman; Faust, 1994; Higgins; Ribeiro; 2018), sendo também uma boa estratégia para investigar conexões entre elementos sociais e culturais. Contudo, ela precisa estar suportada por redes com dados bem estruturados, com conteúdos bem formatados e com integridade, além de contar com o apoio de informações estatísticas para auxiliar na interpretação, o que depende de dados bem descritos.

Assim, o problema de pesquisa encontra-se em saber de que maneira o uso de padrões de documentação consegue favorecer o tratamento de dados endereçados à análise exploratória de dados, de modo a entender suas características, identificar padrões, detectar *outliers* e formular hipóteses preliminares. Por conseguinte, o objetivo geral da pesquisa é identificar agrupamentos coesos nos autores contemplados nos editais por meio de similaridades nos eixos temáticos de seus respectivos projetos, a partir da análise exploratória das bases de dados do Funcultura.

2 PROCEDIMENTOS METODOLÓGICOS

A pesquisa é de natureza aplicada, de abordagem quali-quantitativa, e de caráter exploratório e descritivo, envolvendo bases de dados do acervo cultural do Funcultura, disponibilizado em repositório digital institucional. Essas bases são exploradas e seus dados tratados para a geração de gráficos estatísticos e redes de informação tripartida,

bipartida e monopartida. Os procedimentos metodológicos foram conduzidos em duas etapas, a saber: (i) modelagem e criação das bases de dados; e (ii) análise e exploração dos dados.

A primeira etapa metodológica pode ser considerada a espinha dorsal para a obtenção de dados curados para transformação em fontes de informações qualificadas e destinadas à plataforma de agregação MEDIATECA Capixaba, bem como matéria-prima para gestores da área da cultura em produzir informações estratégicas à tomada de decisões ao Fundo.

Assim sendo, a etapa de Modelagem do Esquema de Metadados foi assistida por meio da técnica de criação de Matriz de Restrições (Baker; Nilsson; Johnston, 2008), em que se integra e prevê um conjunto de registros de metadados que são instâncias válidas para um dado perfil de aplicação, neste caso, o acervo cultural do Funcultura.

A partir do ambiente de desenvolvimento da tecnologia Tainacan³, pôde-se criar o esquema de dados à luz da Matriz de Restrições. Nessa perspectiva, a produção das bases de dados do Funcultura contou com o uso de padrões de documentação (Lemos; Martins; Macedo, 2024) na organização e representação dos documentos do Fundo, destacando o uso de vocabulários controlados (taxonomias e arquivos de autoridade) e do padrão de metadados *Dublin Core* como um modelo de referência, visando obter informações normalizadas e consistentes para a extração de *insights* valiosos à instituição.

Na segunda etapa metodológica, voltada à análise exploratória, os dados oriundos das coleções do acervo do Funcultura, usados na pesquisa, foram organizados em três tabelas relacionadas. A primeira tabela representa as informações dos 876 autores proponentes dos projetos que foram contemplados. A segunda tabela contém as informações dos 1370 projetos inscritos que foram contemplados bem como o edital ao qual participaram, sendo conectada à primeira tabela por meio de um identificador único do autor proponente. A terceira tabela possui informações sobre 262 produtos vinculados aos projetos contemplados, sendo conectada à segunda tabela por meio de um

³ Tainacan é o *software* usado na implementação das bases de dados na rede de provedores da MEDIATECA Capixaba. Disponível em: <https://tainacan.org/>.

identificador único do projeto inscrito. Todos os dados da pesquisa, inclusive os arquivos obtidos nos processamentos dos resultados estão disponíveis em repositório para acesso público⁴.

Com apoio do *software OpenRefine*⁵, os dados das três tabelas passaram por um processo de preparação e de transformação com foco na qualidade e padronização. Essa etapa envolveu ações, tais como a limpeza de dados com símbolos irrelevantes, eliminação de registros em branco, padronização de nomes de variáveis, e a criação de variáveis a partir da extração de dados existentes para representar, com menor granularidade, informações sobre o ano do edital, a categoria cultural e o eixo temático ao qual os projetos se enquadram. Ao todo foram organizados, por meio de um vocabulário controlado do tipo taxonomia, oito categorias culturais e 49 eixos temáticos dentre os 1370 projetos contemplados. A distribuição da quantidade de projetos nessa taxonomia pode ser vista no Gráfico 1, na seção de resultados.

Com foco na identificação de agrupamentos coesos entre os autores contemplados, por meio de similaridades nas categorias culturais e eixos temáticos de seus respectivos projetos, criou-se um mapeamento dos dados das tabelas para o formato de uma rede de informação tripartida contendo os seguintes nós: autores contemplados, editais de inscrição, e eixos temáticos. O referido mapeamento, Quadro 1, foi processado no *software OpenRefine* e gerou um arquivo no formato GML⁶ para ser interpretado pelo *software Gephi*⁷.

No próprio *software Gephi*, a partir da rede tripartida gerada pelo mapeamento do Quadro 1, foram aplicadas projeções bipartidas para a obtenção de outras redes simplificadas, especialmente uma rede monopartida de autores contemplados.

⁴ Repositório de dados da pesquisa. Disponível em:

<https://drive.google.com/drive/folders/1AQUF007OGKcxI1097FmYHGJy0v3Lyeku?usp=sharing>.

⁵ OpenRefine é considerado um *Interactive Data Transformation*, sendo uma ferramenta de código aberto utilizada para a limpeza e transformação de volumes massivos de dados estruturados. Disponível em:

<https://openrefine.org/>.

⁶ *Graph Modelling Language* (GML) é um formato para representação de grafos de fácil leitura por humanos e com uma capacidade semântica razoável para configurar as características da rede, dos nós e das arestas. Disponível em: https://en.wikipedia.org/wiki/Graph_Modelling_Language/.

⁷ Gephi é um *software* de código aberto utilizado para visualização, análise e manipulação de redes e grafos. Disponível em: <https://gephi.org/>.

Quadro 1 - Mapeamento de dados para o formato de rede executado no *software* OpenRefine

```
graph [
comment "Rede de contemplados por eixo e chamada do edital"
directed 0

node [ id {{jsonize(cells.idContemplado.value)}}
categoria "idContemplado"
grupo "contemplado"
eixo {{jsonize(cells.eixo.value)}}
chamada {{jsonize(cells.chamadaEdital.value)}} ]

node [ id {{jsonize(cells.eixo.value)}}
categoria "eixo"
grupo "eixo-chamada" ]

node [ id {{jsonize(cells.chamadaEdital.value)}}
categoria "chamada edital"
grupo "eixo-chamada" ]

edge [ source {{jsonize(cells.idContemplado.value)}} target
{{jsonize(cells.eixo.value)}} ]

edge [ source {{jsonize(cells.idContemplado.value)}} target
{{jsonize(cells.chamadaEdital.value)}} ]
]
```

Fonte: Dados da pesquisa (2024).

O trabalho recorreu à técnica de análise de redes sociais (ARS) para a descoberta de conhecimento e, principalmente, por ser uma ferramenta metodológica que viabiliza enxergar o que outras abordagens não permitem (Wasserman; Faust, 1994; Higgins; Ribeiro, 2018), sendo também uma boa estratégia para investigar conexões entre elementos sociais e culturais. Aliado a ARS, a inspeção visual foi amplamente usada por ser uma ação que direciona a investigação em novas explorações. Contudo, ela precisa estar suportada por redes com uma boa formatação dos dados, além de contar com o apoio de gráficos estatísticos para auxiliar na interpretação. Dessa forma, foram usados os *softwares* Gephi, Looker Studio⁸ e uma planilha eletrônica nessa etapa de análise e interpretação, cujos artefatos visuais e resultados são apresentados na próxima seção.

3 RESULTADOS E DISCUSSÃO

O levantamento de requisitos (Quadro 2) foi fundamental para a organização dos dados do acervo em três coleções (bases de dados): projeto, contemplado (autor) e

⁸ Looker Studio é um *software* gratuito para *data visualization* e criação de relatórios e painéis interativos a partir de dados. Disponível em: <https://lookerstudio.google.com/>.

produto. Cada coleção foi organizada com as suas principais entidades que demandaram metadados para as suas descrições e representatividades nas bases de dados.

Quadro 2 - Especificação de requisitos das principais entidades das bases de dados do Fundo

Entidade	Descrição	Coleção Tainacan
Dados gerais	Fornece informações administrativas sobre os lançamentos anuais dos editais, suas respectivas categorias e eixos.	Funcultura - Projeto
Projeto contemplado	Fornece informações sobre o projeto contemplado, incluindo título, proponente contemplado, o valor recebido, contatos do agente cultural e seu relatório final de prestação de contas.	
Produto [outra coleção a ser criada no Tainacan, e representada via tipo de metadado relacionamento]	Fornece informações sobre as diferentes ações entregues pelo proponente em forma de resultados do projeto contemplado. Logo, um edital pode se desdobrar em vários produtos. O produto, por sua vez, é o objetivo de um edital, ou seja, o resultado esperado pelo proponente ao submeter uma proposta de projeto.	
Contemplado	Pessoas envolvidas com o desenvolvimento do produto participante de um edital e projeto.	Funcultura – Contemplado
Categoria de Metadados	Com título, descrição e tipo de produto (lista de seleção).	Funcultura – Produto
Acervo	Fornece informações sobre um conjunto de objetos, históricos ou não, que necessitam estar reunidos para ocorrer sua compreensão.	
Audiovisual	Fornece informações sobre um produto que necessariamente apresente imagem e som de forma engajada.	
Documento técnico	Fornece informações registradas em papel ou meio eletrônico que possa ser usado para consulta, estudo ou validação técnica.	
Evento	Fornece informações sobre um acontecimento que reúna pessoas, presencialmente ou virtualmente, para determinado fim.	
Livro	Fornece informações sobre uma publicação física ou digital, composto por páginas, que necessariamente deve ter ISBN ou ISSN.	
Mestres da Cultura Popular	Fornece informações biográficas de personalidades da cultura capixaba.	
Música	Fornece informações de uma prática cultural que combine sons e ritmos.	
Site	Fornece informações sobre um sítio eletrônico criado como ou para o produto, acessado através de endereço HTTP ou HTTPS, disponível ou não, na web.	

Fonte: Dados da pesquisa (2024).

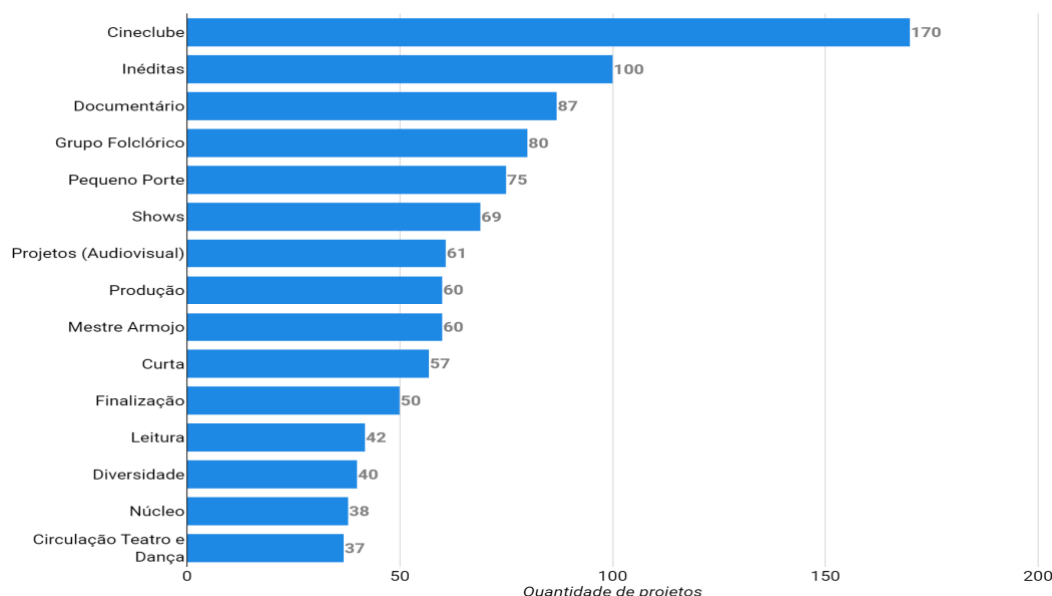
A Matriz de Restrições com a relação dos metadados e seus respectivos perfis de aplicação foi elaborada a partir desses requisitos, evidenciando as decisões de modelagem para cada elemento presente na matriz, tornando-o, assim, uma representação da informação e do conhecimento diferenciada para as bases de dados do Funcultura, em que refletiram em metadados normalizados, padronizados e qualificados à luz dos vocabulários controlados propostos na presente pesquisa.

As taxonomias criadas para o acervo digital do Funcultura são em nove: autor de livro; categoria e eixo; estilo musical; local de atuação dos mestres; local de gravação; localidade; localidade do evento; mestre da cultura popular; e tipo de evento. Para além das taxonomias, vários arquivos de autoridade foram modelados para uso nos valores de dados de alguns metadados, as cardinalidades indicadas, os principais filtros atribuídos, a obrigatoriedade ou não na entrada de dados, as orientações de uso do metadado na perspectiva do usuário, entre outras configurações, garantindo, assim, a qualidade na catalogação dos dados e, conseqüentemente, na recuperação da informação.

Finalmente, uma vez as bases de dados criadas com padrões de documentação fundamentados em princípios teórico-metodológicos da Ciência da Informação, a exemplo da modelagem de taxonomias apropriadas ao contexto de uso, tornou-se possível obter informações relevantes à Instituição a partir de dados qualificados, como pode-se observar nos gráficos e nas redes de informação projetados a seguir.

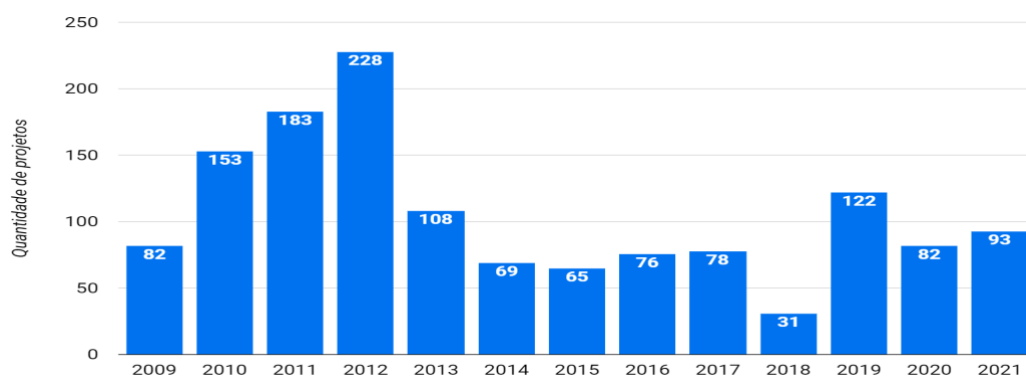
Os resultados de um breve levantamento estatístico sobre a quantidade de projetos distribuídos para os eixos temáticos do Fundo são mostrados no Gráfico 1. Destaca-se também a distribuição da quantidade de projetos ao longo dos anos de publicação dos editais, apresentada no Gráfico 2.

Gráfico 1 - Distribuição dos projetos culturais nos 15 principais eixos temáticos



Fonte: Dados da pesquisa (2024).

Gráfico 2 - Distribuição da quantidade de projetos ao longo dos anos



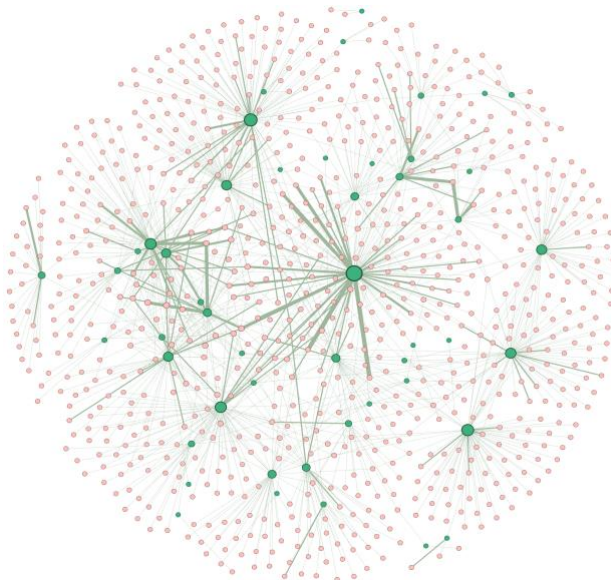
Fonte: Dados da pesquisa (2024).

No que diz respeito às redes de informação, a rede bipartida é composta por nós avermelhados, que representam os autores contemplados, e nós esverdeados, que representam os eixos temáticos, conforme representado na Figura 1.

Essa rede destaca o tamanho dos nós de forma proporcional ao maior grau ponderado. Destaca também a espessura de arestas por aquelas com maior quantidade de conexões coincidentes entre autores com projetos de mesmo eixo temático. Nesse caso, o eixo temático “Cineclube”, na parte central da rede, é destacado, por estar associado a um maior número de autores com projetos contemplados e encontra-se em sintonia com o destaque em ocorrências absolutas que aparece no Gráfico 1. Além disso, devido a múltiplas conexões de muitos nós avermelhados (autores), a rede revela a

existência de um grande número de autores com projetos associados a diversos eixos temáticos.

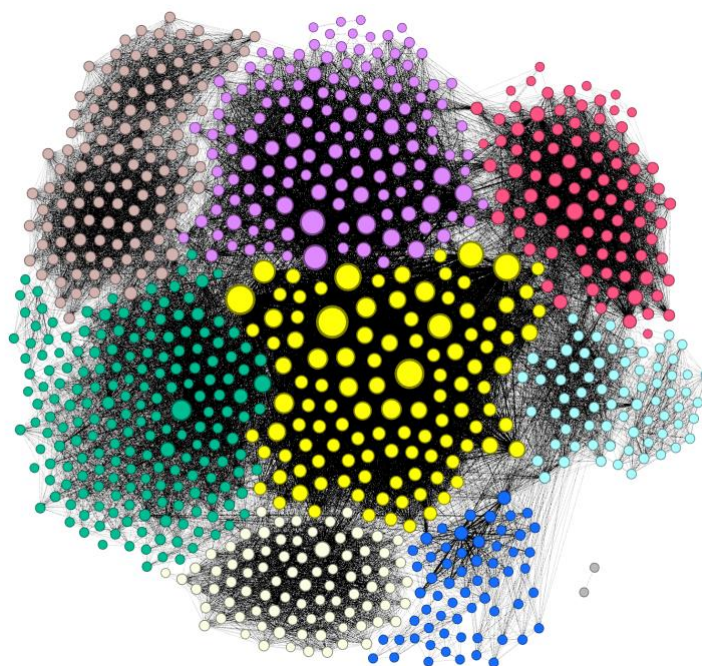
Figura 1 - Rede bipartida com destaque para eixos com maior incidência de inscrições de autores



Fonte: Elaboração própria (2024).

A rede monopartida de autores contemplados, oriunda de uma projeção bipartida, é mostrada na Figura 2.

Figura 2 - Rede monopartida de autores contemplados



Fonte: Elaboração própria (2024).

Para a obtenção dos nove grupos apresentados, destacados em cores diferentes, foi aplicado um algoritmo de clusterização e um algoritmo de distribuição disponíveis no *software* Gephi. O tamanho dos nós foi apresentado de forma proporcional ao grau ponderado dos mesmos.

Em consonância com a rede da Figura 2, a Tabela 1 apresenta a ocorrência dos eixos temáticos nos oito principais grupos de autores contemplados mostrados na Figura 2. Para a obtenção de um resultado mais sistematizado, foram considerados apenas os eixos temáticos cujo percentual de ocorrência de contemplados é maior do que 10% sobre o restante dos eixos temáticos do mesmo grupo.

Os agrupamentos resultantes na rede da Figura 2, analisados por inspeção visual com auxílio da Tabela 1, revelam alguns padrões de incidência de autores que participam regularmente de projetos em eixos temáticos, tais como "Documentário", "Projetos (Audiovisual)", "Curta" e "Finalização", no grupo de cor rosa; "Circulação Teatro e Dança", "Residência Teatro e Dança", "Circo", "Infantil" e "Lona de circo", no grupo cor azul escuro. Por outro lado, o grupo de cor amarela, destaca o eixo "Cineclube" de forma isolada sugerindo uma interpretação de que seus autores (maiores graus ponderados da rede) tendem a permanecer no mesmo eixo.

Tabela 1 - Distribuição dos eixos temáticos nos oito principais grupos da rede da Figura 2

Grupo cor amarela		Grupo cor azul escuro	
Cineclube	96%	Circulação Teatro e Dança	26%
Grupo cor branca		Residência Teatro e Dança	20%
Shows	61%	Circo	18%
Setorial (Literatura)	22%	Infantil	18%
Grupo cor rosa		Lona de circo	15%
Documentário	26%	Grupo cor verde	
Projetos (Audiovisual)	26%	Pequeno Porte	26%
Curta	16%	Núcleo	14%
Finalização	13%	Leitura	12%
Grupo cor bege		Diversidade	12%
Grupo Folclórico	54%	Grupo cor azul claro	

Mestre Armojo	43%
Grupo cor vermelha	
Inéditas	64%

Produção	49%
Bolsa	37%

Fonte: Dados da pesquisa (2024).

A análise exploratória de dados (Castro; Ferrari, 2017) é empregada para examinar, resumir e visualizar os dados para entender suas características principais, identificar padrões, detectar *outliers* e formular hipóteses preliminares, utilizando-se de gráficos, medidas-resumo e análises de distribuição (histogramas, redes de informação etc.) Os Gráficos 1 e 2 e as Redes (Figuras 1 e 2) mostram que a análise exploratória de dados pode ser feita a partir da sumarização realizada em categorias pré-concebidas (com o uso de vocabulário controlado nas bases de dados). Desse modo, é possível perceber que, o que distingue dado ou um conjunto de dados de informação é o conhecimento que ela propicia ao tomador de decisões (Hjørland, 2018).

Logo, na base de dados de Projetos, por exemplo, nota-se que o dado “identificação dos projetos”, representado pelo metadado “ID_Projeto”, é classificado numa categoria oriunda da taxonomia Categoria-Eixo. Por meio desta forma de modelagem de dados, foi possível gerar visualizações de informações importantes acerca das categorias culturais e eixos temáticos envolvidos nos editais, projetos e grupos de contemplados. O gestor de cultura da Instituição poderá, a partir de conhecimentos e habilidades técnicas, tomar decisões a partir das informações geradas frente às bases de dados, incluindo elaboração de estudos do perfil dos contemplados participantes de editais; extração de *insights* valiosos sobre o perfil do público, sua preferência em ações culturais, entre outras. Enfim, essas informações podem ser usadas pelos gestores para melhorar a experiência com os editais e tipos de projetos, além da possibilidade de criar editais personalizados e direcionar melhor recursos públicos e estratégias de captação de produtos e serviços culturais para o Estado.

Do mesmo modo, a partir da relação entre os dados “ano dos projetos” e “identificação dos projetos” (Gráfico 2), tornou-se possível projetar a distribuição dos projetos ao longo dos períodos na trajetória do Funcultura. Assim, o gestor do Fundo também poderá, a partir desse potencial indicador, tomar decisões de investir em

programas e políticas públicas para setores culturais que se sobressaíram nos últimos anos; além de identificar tendências de interesse dos contemplados frente aos produtos desenvolvidos anualmente a partir dos projetos e seus eixos temáticos. Esses *insights* podem orientar o desenvolvimento de curadorias digitais na plataforma Midiateca, com trajetos narrativos temáticos do Fundo que estimulem a participação de artistas capixabas nos editais ofertados pelo governo do Estado.

3 CONSIDERAÇÕES FINAIS

O trabalho apresentou o percurso da construção das bases de dados padronizadas e qualificadas para o Fundo de Cultura e, utilizando-se de análise exploratória, identificou agrupamentos coesos nos autores contemplados nos editais por meio de similaridades nos eixos temáticos de seus respectivos projetos.

O estudo reforçou a importância dos padrões de documentação para qualificar os dados adotando modelagem de metadados e vocabulários controlados, estruturando e padronizando as bases de dados.

A análise de redes de informação revelou relações entre autores, editais e temas, evidenciando padrões de participação e distribuição de recursos. Além disso, a análise exploratória de dados permitiu identificar tendências para subsidiar políticas públicas culturais. Assim, a pesquisa contribui para a gestão da informação cultural em ambientes digitais, indicando caminhos para aprimoramentos na organização de acervos e na democratização do acesso à cultura.

REFERÊNCIAS

- BAKER, Thomas; NILSSON, Mikael; JOHNSTON, Pete. **The singapore framework for dublin core application profiles**. 2008. Disponível em: <http://dublincore.org/documents/singapore-framework/>. Acesso em: 04 mar. 2025.
- CASTRO, Leandro Nunes de; FERRARI, Daniel Gomes. **Introdução à mineração de dados**. São Paulo: Saraiva Educação SA, 2017.
- ESPÍRITO SANTO. Instituto Jones dos Santos Neves. **Análise executiva do FUNCULTURA/ES**. Vitória, ES, 2023. 2. ed. 61 p.; il. tab. (Relatório de Pesquisa).

GILLILAND, Anne J. Setting the Stage. *In*: BACA, Murta. (ed.). **Introduction to metadata**. 3. ed. Los Angeles: Getty Research Institute, 2016. Disponível em: <https://www.getty.edu/publications/intrometadata/setting-the-stage/>. Acesso em: 04 mar. 2025.

HIGGINS, Silvio Salej; RIBEIRO, Antonio Carlos Andrade. **Análise de redes em Ciências Sociais**. Brasília: Enap, 2018. Disponível em: https://repositorio.enap.gov.br/bitstream/1/3337/1/Livro_Analise%20de%20Redes%20em%20Ci%C3%Ancias%20Sociais.pdf. Acesso em: 04 mar. 2025

HJØRLAND, Birger. Data (with big data and database semantics). **Knowledge Organization**, v. 45, n. 8, p. 643-652, 2018. Disponível em: <https://www.isko.org/cyclo/data>. Acesso em: 04 mar. 2025

LEMOS, Daniela Lucas da Silva; MARTINS, Dalton Lopes; MACEDO, Dirceu Flávio. Plataforma de agregação midiateca capixaba: projeto cultura digital nas instituições de memória do Espírito Santo. *In*: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 24., 2024, Vitória. **Anais [...]**. Vitória: ANCIB, 2024. Disponível em: <https://enancib.ancib.org/index.php/enancib/xxivenancib/paper/viewFile/2565/1516>. Acesso em: 04 de mar. 2025.

WASSERMAN, Stanley; FAUST, Katherine. **Social network analysis: methods and applications**. Cambridge, England; New York: Cambridge University Press, 1994.



[4.0 Internacional](https://creativecommons.org/licenses/by/4.0/)

Esta licença permite compartilhamento, remixe, adaptação e criação a partir do trabalho, mesmo para fins comerciais, desde que sejam atribuídos créditos ao(s) autor(es). Conteúdos elaborados por terceiros, citados e referenciados nesta obra não são cobertos pela licença.

Como citar este trabalho:

LEMOS, Daniela Lucas da Silva; CRISTOVÃO, Henrique Monteiro. Bases de dados qualificadas para a descoberta de conhecimentos úteis à gestão de acervos culturais. *In*: WORKSHOP DE INFORMAÇÃO, DADOS E TECNOLOGIA, 8., Marília, SP. **Anais [...]**. Marília, SP: Universidade de Marília, 2025. DOI: <https://doi.org/10.22477/viii.widat.238>.



WIDaT 2025

VIII WORKSHOP DE INFORMAÇÃO,
DADOS E TECNOLOGIA

MARÍLIA • SP

Restauração digital de fotografias com Inteligência Artificial: o papel das cópias autoritárias na validação do documento produzido

Digital restoration of photographs with Artificial Intelligence: the role of authoritative copies in the validation of the document produced

Telma Campanha de Carvalho Madio¹



<http://lattes.cnpq.br/1139786651111231>



<https://orcid.org/0000-0002-7031-2371>

Vinicius Gustavo Franco²



<http://lattes.cnpq.br/0392641492664608>



<https://orcid.org/0009-0002-0783-5150>

Thiago Rafael Martins³



<http://lattes.cnpq.br/9819341422675074>



<https://orcid.org/0009-0006-9476-8775>

Resumo

A aplicação da Inteligência Artificial (IA) tem influenciado práticas e fundamentos em áreas como a Ciência da Informação, Arquivologia e Biblioteconomia. No campo da restauração de fotografias, anteriormente realizada de forma manual, recursos computacionais e softwares gráficos contribuíram para a reformulação dos processos. Recentemente, a utilização de Redes Generativas Adversárias (GAN) tem permitido a reconstrução digital de imagens danificadas, inclusive com base em descrições textuais. Este estudo teve como objetivo analisar os desafios e implicações do uso de IA na restauração digital de fotografias, com ênfase na aplicação do conceito de cópia autoritária por instituições arquivísticas. A metodologia consistiu em levantamento bibliográfico, com análise de publicações acadêmicas, artigos e documentos institucionais sobre preservação digital, autenticidade documental e tecnologias de IA. Os resultados apontam que, embora os sistemas de IA contribuam para a recuperação de informações visuais, persistem questões relacionadas à rastreabilidade das alterações e à preservação da integridade documental. O estudo também identificou que a adoção do conceito de cópia autoritária por parte de algumas instituições pode ser um meio de validar documentos restaurados digitalmente, desde que acompanhada de diretrizes de controle e registro dos processos aplicados. Conclui-se que a integração da IA à restauração digital demanda parâmetros técnicos e normativos que garantam a transparência dos procedimentos, a manutenção da identidade arquivística e a continuidade do acesso aos documentos. A preservação digital, nesse contexto,

1 Doutora, Universidade Estadual Paulista (UNESP), Marília, São Paulo, Brasil. Professora associada, Universidade Estadual Paulista (UNESP), Marília, São Paulo, Brasil. telma.madio@unesp.br.

2 Mestrando, Universidade Estadual Paulista (UNESP), Marília, São Paulo, Brasil. Aluno de pós-graduação, Universidade Estadual Paulista (UNESP), Marília, São Paulo, Brasil. vinicius.g.franco@unesp.br.

3 Mestrando, Universidade Estadual Paulista (UNESP), Marília, São Paulo, Brasil. Aluno de pós-graduação, Universidade Estadual Paulista (UNESP), Marília, São Paulo, Brasil. thiago.rafael@unesp.br.

apresenta-se como instrumento essencial para a gestão de documentos restaurados com o uso de tecnologias baseadas em IA.

Palavras-chave: preservação; fotografia; Inteligência Artificial.

Abstract

The application of Artificial Intelligence (AI) has influenced practices and foundations in areas such as Information Science, Archivology and Librarianship. In the field of photo restoration, which was previously carried out manually, computer resources and graphics software have contributed to the reformulation of processes. Recently, the use of Generative Adversarial Networks (GAN) has enabled the digital reconstruction of damaged images, including those based on textual descriptions. The aim of this study was to analyze the challenges and implications of using AI in the digital restoration of photographs, with an emphasis on the application of the concept of authoritative copying by archival institutions. The methodology consisted of a bibliographic survey, analyzing academic publications, articles and institutional documents on digital preservation, document authenticity and AI technologies. The results show that although AI systems contribute to the retrieval of visual information, there are still issues related to the traceability of changes and the preservation of document integrity. The study also identified that the adoption of the concept of authoritative copying by some institutions can be a means of validating digitally restored documents, as long as it is accompanied by guidelines for controlling and recording the processes applied. The conclusion is that the integration of AI into digital restoration requires technical and regulatory parameters that guarantee the transparency of procedures, the maintenance of archival identity and continued access to documents. In this context, digital preservation is an essential tool for managing documents restored using AI-based technologies.

Keywords: preservation; photography; Artificial Intelligence.

1 INTRODUÇÃO

A utilização da Inteligência Artificial (IA) tem alcançado os mais variados campos do conhecimento, impactando diretamente as bases teóricas e a aplicação de novas abordagens em áreas antes rigidamente definidas. Para Rockembach (2021), a implementação e a crescente adoção de soluções de IA em diversos setores seguem um caminho irreversível, impactando também a Ciência da Informação, os Arquivos e as Bibliotecas. Nesse contexto, observa-se que a relação entre as inteligências humana e artificial pode se desenvolver em duas direções principais: competição ou colaboração.

A restauração de fotografias, por muitos anos, foi considerada uma prática exequível apenas por meios manuais. No entanto, com o avanço da computação e dos softwares gráficos, como o *Photoshop* da *Adobe*, essa prática passou por uma transformação significativa, reduzindo o tempo de trabalho, ampliando as ferramentas disponíveis e melhorando a qualidade dos resultados.

As etapas do processo de restauração digital passaram a incorporar IA do tipo GAN (Redes Generativas Adversárias)⁴, que são capazes de gerar imagens hiper-realistas⁵ a partir de comandos de texto, auxiliando na recuperação de danos, sujidades e imperfeições acumulados ao longo do tempo. Beiguelman (2023a)⁶, afirma que a IA está desfazendo as linhas entre realidade e ficção, tornando cada vez mais complicado validar a autenticidade de fotos e vídeos gerados por essas IA. No entanto, essa inovação levanta um problema fundamental: se o documento físico original não está sendo diretamente alterado, mas sim transposto para o ambiente digital, qual é a natureza do documento final produzido?

Os documentos em questão são documentos arquivísticos, Bellotto (2006) conceitua como um conjunto de documentos reunidos e estruturados de acordo com suas funções e atividades, com o propósito de preservar a memória e fornecer evidências para ações administrativas e registros históricos. Sobre esses documentos recaem diferentes processos e atividades, como a restauração e a preservação, fundamentais para garantir sua integridade, autenticidade e acesso contínuo.

O tema central aqui abordado é justamente o desafio de manter determinadas características essenciais desses documentos arquivísticos, mesmo diante dessas transformações inevitáveis causadas pela IA. Como assegurar que, após intervenções digitais, seja na recuperação imagética de uma fotografia deteriorada, o documento continue sendo reconhecido como autêntico e confiável? A questão ganha ainda mais relevância no contexto da preservação digital, onde a necessidade de adaptação tecnológica pode entrar em conflito com a exigência de manter a identidade arquivística do documento.

Deste modo, existe atualmente a necessidade de se debruçar e trazer novas reflexões sobre a preservação digital, que desafiam os também tradicionais significados

⁴ Goodfellow (2014), demonstra que esse tipo de Redes Generativas Adversariais (GANs) são um modelo de inteligência artificial que consiste em duas redes neurais: um gerador, que cria dados sintéticos, e um discriminador, que avalia se os dados gerados são reais ou falsos. Esse processo competitivo aprimora a qualidade dos dados gerados, tornando-os cada vez mais realistas (tradução nossa).

⁵ Conceito de hiper-realismo é ligado a um movimento artístico que surgiu nos anos 1960, que se caracteriza por reproduzir a realidade de forma muito precisa, como uma fotografia.

⁶ Referência do trabalho publicado na Rapsódia um almanaque de filosofia e arte, com o título: Inteligência artificial como fármakon: a arte algorítmica entre o remédio e o veneno. Rapsódia.

de conservação e restauração. Como apontado por Chen (2001) e Thibodeau (2002) *Apud* Yamaoka; Gauthier (2013, p.78):

Preservar alguma coisa tradicionalmente significa mantê-la inalterada, isto é, o mais próximo possível do seu estado original. Na preservação digital, para o objeto digital permanecer acessível ao longo do tempo, é inevitável fazer alterações no objeto, ou seja, é necessário mudar para não perecer – é o paradoxo da preservação digital.

Para responder a essa questão, aplica-se o conceito de cópia autoritária, conforme definido pelo projeto InterPARES⁷. O conceito confere às instituições o respaldo necessário para atestar que o documento digital resultante é legítimo, pertence à instituição e pode ser considerado confiável para uso e preservação. A cópia autoritária valida a autenticidade do documento restaurado digitalmente, bem como estabelece diretrizes para sua gestão e acesso futuro, garantindo que sua integridade e fidedignidade sejam mantidas ao longo do tempo.

1.1 Restauração Digital

A restauração digital está em desenvolvimento em diversas áreas, especialmente nas Artes Digitais, enquanto na Arquivologia e na Ciência da Informação inicia-se um debate acadêmico sobre seus impactos e implementação. Inspirada na restauração física, essa prática busca recuperar informações da forma mais fiel possível à captura original, utilizando softwares gráficos e inteligência artificial. Não há um único método, mas múltiplas abordagens, variando de aplicativos prontos à modelagem manual. A IA tem se tornado cada vez mais relevante nesse campo, com sistemas baseados em GAN auxiliando na reconstrução de danos, colorização, remoção de rasuras e ampliação de imagens sem perda de qualidade.

Mas a tecnologia GAN, não vem sem problemáticas, como apontado por Beiguelman (2023b)⁸, imagens hiper-realistas, geradas por IA por meio de ferramentas As IA generativas, como *Midjourney* e *DALL·E*, introduzem uma nova narrativa visual mediada por algoritmos opacos, levantando a questão de até que ponto passaremos a

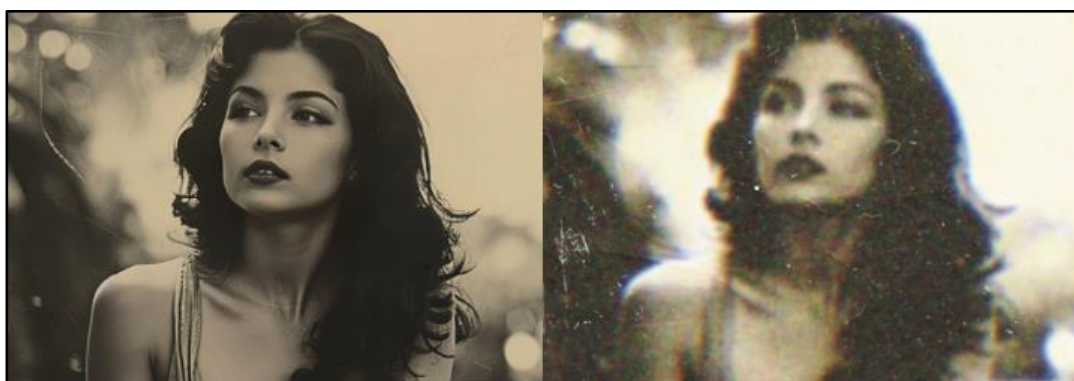
⁷ O Projeto InterPARES 3 é um projeto multidisciplinar internacional que desenvolve práticas e ferramentas para garantir a proteção dos documentos arquivísticos digitais. Disponível em: http://www.interpares.org/ip3/ip3_index.cfm?team=4.

⁸ Referência do trabalho publicado na revista Zum, com o título: Inteligência artificial e as novas políticas das imagens.

enxergar o mundo sob a perspectiva dessas tecnologias. Sua incorporação nos arquivos e unidades de informação é irreversível, exigindo uma compreensão aprofundada de seus impactos, especialmente na transposição do analógico para o digital.

O Arquivo Nacional (2025) possui um vasto acervo iconográfico, cuja restauração digital pode agilizar o tratamento e a disponibilização ao público. No entanto, essa prática não se limita à recuperação de imagens danificadas, mas também melhora a qualidade visual, tornando-as mais acessíveis para a sociedade.

Figura 1 – Captura de tela do site *Airbrush*⁹, o antes e depois de uma restauração por IA.



Fonte: Airbrush (2025).

De fato, as grandes questões postas, quando é realizada a restauração digital é um documento fotográfico, transportando-o de um suporte físico, via digitalização, para um arquivo digital binário, são como nomear, classificar e trabalhar com esse novo documento gerado. Visto que, por ser digital, esse documento será amplamente circulado e utilizado em meios digitais, vale garantir que, mesmo após as inúmeras alterações no processo de restauração, incluindo interferências da IA, ele permaneça confiável, autêntico para o público.

A distinção entre documentos digitais restaurados com inteligência artificial e documentos digitais nativamente produzidos é substancial para o tratamento arquivístico. Os primeiros derivam de suportes físicos e passam por processos de digitalização e intervenção técnica, como correções por softwares gráficos ou redes neurais, gerando versões visuais aprimoradas do original. Já os documentos nativos digitais são criados inteiramente em ambiente digital, sem vínculos diretos com suportes

⁹ O *Airbrush* é uma Inteligência Artificial que possui ferramentas para aprimorar e restaurar imagens. A plataforma pode ser acessada em: <https://airbrush.com/pt/photo-restoration>.

físicos. A diferença implica abordagens distintas quanto à descrição, classificação e preservação. Documentos restaurados com IA exigem registro claro dos métodos aplicados, assegurando transparência e permitindo a comparação com os originais, de modo a manter sua integridade e autenticidade como cópias informacionais.

Em um caso célebre, a fotografia do Santos Dumont¹⁰, sob guarda do Museu Paulista, passou por processos de restauro físico e também sofreu interferências digitais. Segundo Moreira *et al.* (2003), o Museu mantém o original preservado enquanto expõe uma reprodução digitalizada, para difusão, que conserva características tonais semelhantes ao seu estado atual de envelhecimento. Neste contexto, a fotografia analógica resulta de um processo mecânico-físico-químico, sujeita a envelhecimento e danos específicos, enquanto a digital é um arquivo eletrônico gerado por escâneres que convertem a luz refletida em *pixels*¹¹.

Os *pixels*, contendo informações de cor e luminosidade, podem ser manipulados para corrigir danos como rasgos, manchas e oxidação, transformando problemas estruturais em alterações cromáticas. A recuperação digital, ao contrário do retoque convencional que apenas neutraliza contrastes, permite uma reconstrução mais detalhada do original, como demonstrado na restauração do retrato de Santos Dumont.

Assim, compreender a hibridização, além de esclarecer conceitos, evidencia novas possibilidades que estarão presentes nos acervos arquivísticos daqui em diante. Quando uma unidade de informação realiza esse processo, é necessário estabelecer parâmetros para a correta identificação e classificação dos documentos, tanto internamente quanto para seus usuários. Isso requer a explicitação dos métodos utilizados em sua criação, garantindo transparência e assegurando que, por se tratar de uma cópia, o documento original possa ser confrontado com a versão restaurada, reconhecendo-se as diferenças significativas entre ambos.

¹⁰ O processo detalhado do restauro feito pelo Museu Paulista pode ser acessado em: <https://revistas.usp.br/anaismp/article/view/5365/6895>.

¹¹ A definição proposta por Debray (2000), o pixel é a menor unidade de uma imagem digital, resultante da discretização de elementos visuais em pontos individuais. No ambiente digital, ele converte imagens e textos em uma estrutura binária, substituindo a continuidade das linguagens analógicas por sequências de zeros e uns.

1.1.1 Cópias Autoritárias

A adoção do conceito de cópia autoritária não se trata apenas de uma questão semântica ou terminológica para rotular o documento produzido nesta atividade. Pelo contrário, sua aplicação é base para compreender a natureza do documento restaurado, seu contexto de inserção e a forma adequada de tratá-lo arquivisticamente. Os conceitos clássicos de fotografia não podem ser automaticamente aplicados para classificar o resultado de um processo de restauração digital, especialmente quando há interferência de Inteligência Artificial.

Se o que está sendo produzido não é uma mera cópia, pois apresenta alterações significativas em seu conteúdo visual, e tampouco pode ser considerado um documento original ou matriz, visto que apenas se assemelha ao seu referente, surge, então, um dilema conceitual. Para resolvê-lo, aplica-se o conceito de cópia autoritária, desenvolvido pelo projeto InterPARES (2017), manifestação de um documento arquivístico considerada pelo produtor seu documento oficial, geralmente sujeita a procedimentos de controle que não são exigidos em outras manifestações. A restauração digital opera num paradigma tecnológico que exige reflexões aprofundadas sobre autenticidade e integridade documental.

Corroborando com isto Grácio, Fadel e Valentim (2013), a popularização da internet e os avanços contínuos nas Tecnologias de Informação e Comunicação (TIC), a sociedade tem vivenciado transformações significativas na forma de registrar e acessar informações. Antes predominantemente armazenadas em papel, as informações passaram a ser produzidas e preservadas em formato digital.

Nesse contexto, os conceitos e definições estabelecidos pelo InterPARES (2013) são fundamentais, não apenas para embasar a fundamentação teórica e conceitual dessa prática, mas também para garantir a confiabilidade do produto final da restauração digital, ou seja, a cópia autoritária, que preserva os atributos essenciais do documento e atesta sua veracidade.

Além disso, a cópia carrega o aval institucional, propiciando seu uso como fonte legítima de informação em diferentes contextos. A fotografia, enquanto documento arquivístico, demanda a atuação de profissionais da Arquivologia ao longo de todo o seu

ciclo de vida institucional. A restauração, nesse sentido, torna-se uma atividade indispensável, uma vez que os documentos são produzidos em suportes suscetíveis à degradação.

Quando uma fotografia sofre danos ou apresenta perda de qualidade, sua recuperação digital pode ser a única alternativa viável para preservar seu conteúdo informacional. Em restauração de manuscritos, Hanif *et al.* (2023), colocam a restauração digital como uma alternativa viável, pois evita o manuseio excessivo dos documentos e também aborda o resultado alcançados com os avanços feito pela área.

Para que o processo ocorra de maneira criteriosa, é importante que os profissionais da área dominem as ferramentas digitais disponíveis e acompanhem as transformações tecnológicas, mesmo diante de desafios e resistências. O avanço digital é uma realidade consolidada, e sua incorporação aos processos arquivísticos é não apenas necessária, mas inevitável.

A consolidação da restauração digital como prática arquivística exige a adoção de ferramentas tecnológicas avançadas, como também a formulação de diretrizes que assegurem a gestão eficiente dos documentos no ambiente digital. A distinção entre o documento original e sua cópia restaurada deve ser claramente estabelecida, atestando que o registro da intervenção seja transparente e acessível, preservando a autenticidade do material recuperado.

Nesse sentido, a criação de metadados¹² detalhados, a documentação dos processos de restauração e a incorporação de mecanismos de certificação institucional são estratégias necessárias para assegurar que a cópia autoritária mantenha seu valor informacional e jurídico. Além disso, a crescente utilização de IA no aprimoramento da qualidade das imagens reforça a necessidade de reflexões críticas sobre os impactos dessas tecnologias na preservação documental.

O alinhamento entre os avanços tecnológicos e os princípios arquivísticos deve ser constantemente revisado, de modo a garantir que a restauração digital não comprometa a integridade dos documentos, mas sim contribua para sua permanência e acesso no

¹² Para Arakaki e Arakaki (2025, p. 25), metadados são informações estruturadas que descrevem, identificam, organizam ou facilitam o acesso a dados, recursos ou documentos, permitindo sua descoberta, gestão e uso eficiente. Em outras palavras, fornecendo contexto e atributos para tornar a informação mais compreensível e utilizável.

longo prazo. Vale ressaltar que essas reflexões são altamente pertinentes, pois documentos restaurados digitalmente já fazem parte da realidade das unidades de informação.

A falta de diretrizes claras sobre sua organização e preservação pode resultar na perda ou no apagamento dessas informações, comprometendo o acesso e a memória institucional. Assim, a definição de parâmetros arquivísticos específicos para esses documentos é essencial para garantir sua gestão adequada e sua permanência no acervo institucional.

2 PROCEDIMENTOS METODOLÓGICOS

A presente pesquisa adotou a metodologia de levantamento bibliográfico, cujo objetivo foi identificar, selecionar e analisar fontes teóricas e documentais relevantes para a compreensão a influência da IA na restauração digital no contexto arquivístico. O estudo foi desenvolvido com base em obras acadêmicas, artigos científicos, documentos institucionais e glossários que abordam os seguintes eixos temáticos: (i) preservação e restauração de documentos fotográficos; (ii) digitalização e preservação digital; (iii) autenticidade e integridade documental na era digital; e (iv) conceitos arquivísticos, com ênfase na cópia autoritária e no projeto InterPARES.

1- O levantamento bibliográfico foi realizado por meio de pesquisa em bases de dados acadêmicas: Brapci e Google Acadêmico.

2- Sobre os termos utilizados em inglês e português: "restauração digital de fotografias", "preservação digital e autenticidade", "cópia autoritária" e "inteligência artificial e recuperação de imagens". Inicialmente, foram encontrados aproximadamente: 722 o número total de resultados encontrados documentos. Após aplicação do critério qualitativo e temporal, permaneceram 52 resultados. Com a leitura integral e aplicação dos critérios de relevância, foram selecionados 16 textos para compor o *corpus* teóricos da pesquisa.

3- A apuração, para garantir a relevância das fontes, foram adotados os seguintes critérios de inclusão e exclusão.

Inclusão: Trabalhos publicados nos últimos 20 anos (salvo obras clássicas da área); artigos revisados por pares, dissertações, teses e livros acadêmicos; publicações de fontes oficiais e normativas arquivísticas; pesquisas que tratem diretamente da relação entre IA, preservação digital e autenticidade documental.

Exclusão: Materiais de fontes não verificadas ou sem embasamento teórico; trabalhos que abordem apenas aspectos subjetivos e artísticos de documentos imagéticos, sem relação com princípios arquivísticos ou com IA.

4- Após a seleção das referências, foi feita uma análise crítica do conteúdo, categorizando os textos conforme os temas abordados. O material coletado foi organizado em fichamentos, destacando conceitos-chave, abordagens metodológicas e principais contribuições dos autores. A sistematização das informações embasou a discussão teórica do estudo, correlacionando os aspectos técnicos da restauração digital sob os impactos da IAs e as diretrizes do projeto InterPARES.

3 RESULTADOS

Os resultados da pesquisa, a partir de contribuições teóricas de Beiguelman (2023a), Rockenbach (2021), Correadeira (2023) e Hanif *et al.* (2023) indicam que a utilização da IA na restauração digital de fotografias arquivísticas apresenta benefícios significativos, como a automação de processos, a melhoria na qualidade das imagens recuperadas e a redução do tempo necessário para a restauração. As GAN, por exemplo, demonstraram eficácia na reconstrução de áreas danificadas sem comprometer a legibilidade ou a estética do documento original. Entretanto, a pesquisa também revelou desafios críticos relacionados à autenticidade e à confiabilidade dos documentos restaurados digitalmente.

A ausência de padronização na aplicação da IA pode gerar variações nos resultados, dificultando a rastreabilidade das modificações feitas no documento digital. Além disso, a falta de diretrizes claras sobre a preservação de versões anteriores das imagens restauradas pode comprometer a credibilidade dos documentos ao longo do tempo. A adoção do conceito de cópia autoritária mostrou-se uma solução viável para atestar a autenticidade dos documentos restaurados digitalmente. Instituições

arquivísticas que seguem os princípios do projeto InterPARES conseguem estabelecer protocolos rigorosos para a validação, descrição e preservação das imagens restauradas, garantindo sua confiabilidade para futuras consultas e pesquisas.

Dessa forma, a pesquisa conclui que a IA é uma ferramenta valiosa para a restauração digital de fotografias arquivísticas, desde que sua aplicação seja acompanhada de políticas e diretrizes que assegurem a autenticidade e a integridade dos documentos restaurados. A implementação de normativas institucionais e o desenvolvimento de parâmetros específicos para a gestão desses documentos são essenciais para garantir a sustentabilidade da preservação digital no contexto arquivístico.

4 CONSIDERAÇÕES FINAIS

A pesquisa destacou que a Inteligência Artificial é um avanço na restauração digital de fotografias arquivísticas, melhorando a recuperação de imagens degradadas e otimizando a preservação. Contudo, sua aplicação requer cuidado, especialmente quanto à autenticidade e integridade dos documentos restaurados.

O conceito de cópia autoritária é crucial para validar a confiabilidade dos documentos digitais, e seguir normas de instituições como o InterPARES é essencial para preservar a identidade arquivística. Conclui-se que a IA deve ser aplicada de forma equilibrada, com protocolos claros, capacitação profissional e revisão constante das práticas adotadas para garantir a preservação segura do patrimônio fotográfico.

A ausência de padrões e diretrizes específicas pode comprometer a integridade dos registros ao longo do tempo. Nesse cenário, destaca-se a necessidade de implementar normas institucionais, protocolos claros e estratégias que garantam a preservação das versões originais e a rastreabilidade das alterações. Conclui-se que, para que a aplicação da IA seja segura e eficaz no contexto arquivístico, é fundamental adotar uma abordagem equilibrada, com planejamento técnico, políticas de preservação e capacitação contínua, assegurando a integridade e o valor dos documentos restaurados para a sociedade.

REFERÊNCIA

- ARAKAKI, Ana Carolina Simionato; ARAKAKI, Felipe Augusto. Dados e metadados: conceitos e relações. **Ciência da Informação**, Brasília, DF, v. 49, n. 3, p. 34-45, 2020. Disponível em: <https://repositorio.unesp.br/handle/11449/180490>. Acesso em: 10 fev. 2025.
- BEIGUELMAN, G. Inteligência artificial como phármakon: a arte algorítmica entre o remédio e o veneno. **Rapsódia**, São Paulo, v. 1, n. 17, p. 161-174, 2023a. Disponível em: <https://doi.org/10.11606/issn.2447-9772.i17p161-174>. Acesso em: 11 fev. 2025.
- BEIGUELMAN, G. Inteligência artificial e as novas políticas das imagens. **Revista Zum**, São Paulo, 5 abr. 2023a. Disponível em: <https://revistazum.com.br/colunistas/inteligencia-artificial-e-as-novas-politicas-das-imagens/>. Acesso em: 9 set. 2024.
- BELLOTTTO, Heloísa Liberalli. **Arquivos permanentes**: tratamento documental. 3. ed. Rio de Janeiro: FGV, 2006. 318 p.
- BRASIL. Arquivo Nacional. **DIBRARQ**. Disponível em: <https://dibrarq.arquivonacional.gov.br/>. Acesso em: 10 abr. 2025.
- CHEN, Su-Shing. Paradox of digital preservation. **Computer**, v. 34, n. 3, p. 24-28, abr. 2001. Disponível em: https://www.researchgate.net/publication/2955408_Paradox_of_digital_preservation. Acesso em: 10 abr. 2025.
- CORREDERA, J. R. Casar. Inteligencia Artificial Generativa. **Anales de la Real Academia de Doctores**, v. 8, n. 3, 2023. Disponível em: <https://www.rade.es/imageslib/PUBLICACIONES/ARTICULOS/V8N3%20-%2001%20-%20ED%20-%20CASAR.pdf>. Acesso em: 20 fev. 2025.
- DEBRAY, Régis. Les révolutions médiologiques dans l'Histoire. **Bulletin des bibliothèques de france**, v. 45, n. 1, p. 4-12, 2000. Disponível em: http://bbf.enssib.fr/bbf/html/2000_45_1/2000-1-p4-debray.xml.asp. Acesso em: 25 fev. 2025.
- GOODFELLOW, Ian *et al.* Generative adversarial nets. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, 27., 2014, Montreal. **Proceedings [...]**. Montreal: Curran Associates, 2014. Disponível em: https://proceedings.neurips.cc/paper_files/paper/2014/file/f033ed80deb0234979a61f95710dbe25-Paper.pdf. Acesso em: 10 abr. 2025.
- GRÁCIO, José Carlos Abbud; FADEL, Bárbara; VALENTIM, Marta Lígia Pomim. Preservação digital nas instituições de ensino superior: aspectos organizacionais, legais e técnicos. **Perspectivas em Ciência da Informação**, v. 18, p. 111-129, 2013. Disponível em:

<https://www.scielo.br/j/pci/a/XnvBfYVhjpzxWPQ79NwFCb/?format=html&lang=pt>.

Acesso em: 10 fev. 2025.

HANIF, M. *et al.* Restoration and content analysis of ancient manuscripts via color space-based segmentation. **Plos One**, v. 18, n. 3, 2023.

INTERPARES. **Cópia autoritária**. 2007. Disponível em:

http://www.interpares.org/ip3/ip3_terminology_db.cfm?letter=d&term=1051. Acesso em: 13 fev. 2025.

MOREIRA, Y. L. P. M. *et al.* Restauração da fotopintura Retrato em tamanho natural de Santos Dumont, do fotógrafo Giovanni Sarracino. **Anais do Museu Paulista**, v. 6-7, p. 161-183, 2003. Disponível em: <https://www.redalyc.org/articulo.oa?id=27300708>. Acesso em: 20 fev. 2025.

ROCKEMBACH, M. Ciência da informação e inteligência artificial: um caminho para arquivos e bibliotecas inteligentes. In: SILVA, Carlos Guardado da; REVEZ, Jorge; CORUJO, Luís. **Organização do Conhecimento no Horizonte 2030: desenvolvimento sustentável e saúde**. Lisboa: Universidade de Lisboa, 2021. Disponível em: <https://lume.ufrgs.br/bitstream/handle/10183/233477/001134518.pdf?sequence=1>. Acesso em: 11 fev. 2025.

THIBODEAU, Kenneth. **Overview of technological approaches to digital Preservation and challenges in coming years**. 2002. Disponível em:

<https://www.clir.org/pubs/reports/pub107/thibodeau/>. Acesso em: 10 fev. 2025.

YAMAOKA, Eloi Juniti; GAUTHIER, Fernando Ostuni. Objetos digitais: em busca da precisão conceitual; objetos digitais. **Informação & Informação**, Londrina, v. 18, n. 2, p. 77-97, ago. 2013. Universidade Estadual de Londrina. Disponível em: <https://ojs.uel.br/revistas/uel/index.php/informacao/article/view/16162>. Acesso em: 10 fev. 2025.



[4.0 Internacional](https://creativecommons.org/licenses/by/4.0/)

Esta licença permite compartilhamento, remixe, adaptação e criação a partir do trabalho, mesmo para fins comerciais, desde que sejam atribuídos créditos ao(s) autor(es). Conteúdos elaborados por terceiros, citados e referenciados nesta obra não são cobertos pela licença.

Como citar este trabalho:

MADIO, Telma Campanha de Carvalho; FRANCO, Vinícius Gustavo; MARTINS, Thiago Rafael. Restauração digital de fotografias com Inteligência Artificial: o papel das cópias autoritárias na validação do documento produzido. In: WORKSHOP DE INFORMAÇÃO, DADOS E TECNOLOGIA, 8., Marília, SP. **Anais [...]**. Marília, SP: Universidade de Marília, 2025. DOI: <https://doi.org/10.22477/viii.widat.249>.



WIDaT 2025

VIII WORKSHOP DE INFORMAÇÃO,
DADOS E TECNOLOGIA

MARÍLIA • SP

Reuso de dados abertos culturais na Governança de Dados em Cidades Inteligentes: estudo com dados do Arquivo Nacional

*Reuse of cultural open data in Data Governance in Smart Cities:
study with data from the National Archive*

Luiz Carlos da Silva¹



<http://lattes.cnpq.br/2553964864422495>



<https://orcid.org/0000-0002-1443-5879>

Luciana Itida Ferrari²



<http://lattes.cnpq.br/8738517383983971>



<https://orcid.org/0000-0002-8223-039X>

Resumo

A Governança de Dados é fundamental para o desenvolvimento e as ações nas Cidades Inteligentes, porque elas utilizam tecnologias e a integração de dados para otimizar a gestão dos serviços públicos e privados, com objetivos de melhorar a mobilidade urbana, reduzir o consumo de recursos e promover o bem-estar social com ações imediatas e preditivas. No entanto, observa-se que dados culturais, especificamente dados de patrimônio cultural, são subutilizados no contexto de Cidades Inteligentes. O objetivo geral deste trabalho é demonstrar através de planilha oriunda do Arquivo Nacional, o potencial de reutilização de dados abertos, como ferramenta para a Governança de Dados. De forma específica, este estudo busca: Apresentar os conceitos, princípios e a importância da Governança de Dados no contexto de Cidades Inteligentes; demonstrar como esses dados podem ser reaproveitados para disseminar conhecimento de maneira eficiente, utilizando *dashboards* interativos. Foi desenvolvido um *dashboard* no Power BI apresentando os dados. Conclui-se que a Governança de Dados, aliada ao uso de dados culturais e à tecnologia, oferece um caminho promissor para o desenvolvimento das Cidades Inteligentes, permitindo que a cultura e a história sejam não apenas preservadas, mas também exploradas e reutilizadas em prol da sociedade.

Palavras-chave: governança de dados; cidades inteligentes; reuso de dados; patrimônio cultural.

Abstract

Data governance is essential for the development and actions in Smart Cities, because they use technologies and data integration to optimize the management of public and private services, with the aim of improving urban mobility, reducing resource consumption and promoting social well-being with immediate and predictive actions. However, it is observed that cultural data, specifically cultural heritage data, are underutilized in the context of Smart Cities. The general objective of this work is to demonstrate, through a spreadsheet from the National Archives, the potential for reusing open data as a tool for data governance.

¹ Doutor em Ciência da Informação pela Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, MG, Brasil. Docente da Universidade Federal do Espírito Santo (UFES), Vitória, ES, Brasil. Luiz.c.silva@ufes.br.

² Doutorado em Engenharia de Sistemas e Computação pela Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro, RJ, Brasil. Docente da Universidade Federal do Espírito Santo (UFES), Vitória, ES, Brasil. Luciana.Ferrari@ufes.br.

Specifically, this study seeks to: Present the concepts, principles and importance of data governance in the context of Smart Cities; demonstrate how this data can be reused to disseminate knowledge efficiently, using interactive dashboards. A dashboard was developed in Power BI presenting the data. It is concluded that data governance, combined with the use of cultural data and technology, offers a promising path for the development of Smart Cities, allowing culture and history to be not only preserved, but also explored and reused for the benefit of society.

Keywords: data governance; smart cities; data reuse; cultural heritage.

1 INTRODUÇÃO

As Cidades Inteligentes, ou *Smart Cities*, vem ganhando destaque como uma resposta às demandas por eficiência, sustentabilidade e qualidade de vida nas áreas urbanas. As cidades utilizam tecnologias e a integração de dados para otimizar a gestão dos serviços públicos e privados, com objetivo de melhorar a mobilidade urbana, reduzir o consumo de recursos e promover o bem-estar social.

Nas Cidades Inteligentes, onde a quantidade de dados é gerada continuamente por sensores, câmeras, dispositivos móveis e outros sistemas, a Governança de Dados (GD) garante que essas informações sejam usadas para criar soluções que beneficiam a sociedade, através da criação de políticas, processos e estruturas para garantir que os dados sejam coletados, armazenados, processados, utilizados e reutilizados de forma imediata ou preditiva.

As soluções através do reuso de dados para fins sociais podem garantir a melhoria dos serviços de saúde, educação, promoção da inclusão social e proteção do meio ambiente. Uma das áreas de grande potencial de utilização dos dados é a preservação e difusão do patrimônio cultural, especialmente através do uso de documentos arquivísticos. Esses documentos podem ser: registros históricos, mapas, fotografias e outros. São fontes valiosas de informação sobre a história, memória e a cultura de uma cidade. Podem ser digitalizados, organizados e integrados em sistemas de dados urbanos, para fins científicos, educacionais, turísticos e de preservação da memória social.

Por isso, podemos pensar o reuso de dados provenientes de planilhas de inventários de documentos como uma prática que ilustra esse potencial. Essas planilhas, que documentam o processo de inventário de arquivos, podem ser reutilizadas para criar

outros projetos culturais, como: exposições virtuais, aplicativos educacionais ou programas para inclusão social.

Vale destacar que a intenção primordial desta pesquisa era realizar a análise empírica dos dados abertos culturais do Fundo de Cultura da Lei João Bananeira (Cariacica-ES). No entanto, devido à ausência de dados disponíveis, foi necessário adaptar o estudo utilizando uma planilha do Arquivo Nacional, que contém dados de inventários de documentos.

Neste contexto, apresentamos como a Governança de Dados (GD), aplicada em Cidades Inteligentes, pode ser uma ferramenta que auxilia a promoção social, especialmente na preservação da memória e difusão do patrimônio cultural. Ao explorar um caso prático de reutilização de dados arquivísticos, onde as informações foram processadas e apresentadas através de *dashboard* interativo, este trabalho pretende demonstrar as oportunidades e desafios de integrar dados culturais no contexto digital.

Constatamos, conforme estudos realizados por Castro e Baracho (2020) apontam que o termo cultura é pouco pesquisado, quando associado a Cidades Inteligentes. Em suas pesquisas, o termo patrimônio cultural é citado como subtópico tal qual a cultura, conduzindo a percepção de pouca importância e relevância. O reuso de dados provenientes de projetos culturais é subaproveitado, apesar de ser importante para a identidade urbana. Por isso, o objetivo geral deste trabalho é demonstrar através de planilha oriunda do Arquivo Nacional, o potencial de reutilização de dados abertos, como ferramenta para a Governança de Dados. De forma específica, este estudo busca: Apresentar os conceitos, princípios e a importância da Governança de Dados no contexto de Cidades Inteligentes; demonstrar como esses dados podem ser reaproveitados para disseminar conhecimento de maneira eficiente, utilizando *dashboards* interativos.

2 GOVERNANÇA DE DADOS E CIDADES INTELIGENTES

Na pesquisa a seguir, serão fundamentados conceitos essenciais (sem a pretensão de esgotar o conhecimento) de Governança de Dados, definida como um conjunto de práticas que asseguram a qualidade e a segurança dos dados em uma

instituição. Abordaremos também as Cidades Inteligentes, caracterizadas pelo uso de tecnologias digitais para melhorar a vida dos habitantes e otimizar recursos.

A GD é um sistema que envolve a criação de processos estruturados para a tomada de decisões, bem como a atribuição clara de responsabilidades para a gestão de todas as atividades relacionadas aos dados. Esse processo é fundamentado em políticas, normas e restrições para garantir não apenas a conformidade regulatória, mas a integridade, segurança e a qualidade dos dados em toda a instituição.

A instituição *The Data Governance Institute* (DGI), através de suas expertises afirma que:

Todas as organizações precisam ser capazes de tomar decisões sobre como gerenciar dados, obter valor deles, minimizar custos e complexidade, gerenciar riscos e garantir a conformidade com requisitos legais, regulatórios e outros requisitos cada vez maiores. [...] Precisam chegar a um consenso sobre como "decidir como decidir". Criar regras, garantir que as regras estejam sendo seguidas e lidar com não conformidades, ambiguidades e problemas. Em suma, necessitam fazer mais do que gerenciar dados; precisam de um sistema de governança que defina as regras de engajamento para as atividades de gerenciamento (The Data Governance Institute, 2024, *online*, tradução nossa)³.

A GD emergiu como um tema central nas discussões científicas refletindo a complexidade e volume dos dados gerados e utilizados. À medida que as instituições reconhecem o valor estratégico dos dados, a necessidade de implementar práticas eficazes de GD torna-se essencial para a transformação desses dados em insights. Diante a assertiva nos relata Barata que:

A GD tem como responsabilidade o controle e gerenciamento dos dados da organização, sendo a responsável por todo controle dos dados e a transformação deles em informações. A GD é um tema novo que vem ganhando destaque e importância nas organizações, principalmente em organizações onde existe uma grande massa de dados e estes são de extrema importância para a tomada de decisões (Barata, 2015, p. 32).

³ Original: "All organizations need to be able to make decisions about how to manage data, realize value from it, minimize cost and complexity, manage risk, and ensure compliance with ever-growing legal, regulatory, and other requirements. [...] They need to reach consensus on how to "decide how to decide." They need to create rules, ensure that the rules are being followed, and to deal with noncompliance, ambiguities, and issues. In short, they need to do more than manage data; they need a governance system that sets the rules of engagement for management activities" (The Data Governance Institute, 2024, *online*).

O pesquisador Rêgo entende que “Governança de Dados é o exercício de autoridade e controle (planejamento, monitoramento e engajamento) sobre o gerenciamento de ativos de dados [...]” (Rêgo, 2013, p. 34). Portanto, essa definição evidencia que a Governança de Dados não é apenas uma questão técnica, mas envolve também a articulação de políticas e práticas que asseguram a qualidade, segurança e usabilidade dos dados ao longo de seu ciclo de vida. Comunica também que a “Governança de Dados é de alto nível, ou seja, é gestão estratégica de dados na esfera executiva” (Rêgo, 2013, p. 12). Isso implica que a Governança de Dados não só estabelece as diretrizes e políticas, mas também alinha a gestão de dados aos objetivos e metas estratégicas da organização.

Quanto ao conceito de Cidades Inteligentes e sustentáveis, este surgiu como um foco significativo nas estratégias de desenvolvimento urbano no mundo. Com o avanço das tecnologias, as cidades aproveitam soluções inovadoras para enfrentar os desafios urbanos, e criarem ambientes mais eficientes, habitáveis e ecologicamente corretos.

Verificamos que as Cidades Inteligentes e Humanas partiram do contexto a seguir:

As primeiras iniciativas para reconfiguração do espaço urbano tiveram origem na reocupação/reurbanização de áreas degradadas, em geral abandonadas pelas “velhas” atividades econômicas, com a instalação de atividades relacionadas à economia criativa, ou seja, atividades culturais e criativas dinâmicas, associadas a jovens profissionais altamente qualificados e multiculturais. Dentre estas atividades, estariam moda, design, software, indústria de filmes, biotecnologia, microeletrônica etc. (Mendes, 2020, p. 5)

O paradigma da cidade inteligente, desenvolvido por atores sociais, destaca a capacidade das cidades de valorizar seus ativos urbanos por meio de soluções inovadoras. O objetivo é melhorar a qualidade de vida, impulsionar a gestão urbana, promover impacto positivo no meio ambiente, fomentar o crescimento econômico e aprimorar infraestrutura e serviços.

Para definir a cidade inteligente citamos a autora Benites, porque ela engloba essas questões apontadas e outros referentes às tecnologias para solução dos problemas quanto às desigualdades, sustentabilidade e resiliência:

Isto posto, um exemplo de solução de *Smart City* de alto nível de inteligência ofereceria serviços fundamentados em algoritmos analíticos sobre nuvens de

dados para redução de congestionamentos e melhoria do transporte público, para moderar o consumo de água e regularizar seu armazenamento e distribuição, para planejar a **logística a eventos culturais**, etc. E liberaria as bases de dados ao acesso dos cidadãos, criando canais (também virtuais) para que eles possam colaborar com o conteúdo e influenciar no desenho dos serviços públicos e políticas, participando mais intensamente do dia-a-dia da cidade e de sua resiliência (Benites, 2016, p. 33).

Importante também ressaltar Bouskela (2016, p. 33, grifo nosso), que enfatiza: “Em uma *Smart City*, a tecnologia conecta cidadãos e empresas à cidade e entre si, eliminando as ilhas de informação e reduzindo impactos negativos com a distribuição inteligente dos recursos”.

Sugerimos que o modelo ideal de implantação em Cidades Inteligentes para combater desigualdades combine as abordagens “*top-down*”⁴ e “*bottom-up*”⁵, beneficiando assim uma parcela maior da sociedade. Defendemos uma participação coletiva na definição dos rumos e no desenvolvimento das cidades, com melhorias que as tornem mais aprazíveis. Esse processo deve envolver tanto as autoridades quanto os cidadãos comuns, facilitando a integração da cultura e promovendo a representação dos patrimônios culturais nas decisões que moldam as Cidades Inteligentes. O uso de dados abertos pode beneficiar diretamente nesse processo.

A estratégia Nacional de Governo Digital, instituída na Lei nº 14.129, de 29 de março de 2021 (Brasil, 2021), menciona diretamente a promoção e uso de dados abertos, que por sua vez são regulados no âmbito do Governo Federal pelo Decreto nº 8.777 de 11 de maio de 2016 (Brasil, 2016). Nesse contexto, encontram-se dados abertos de diversos órgãos disponíveis para consulta pública. Dentre esses, o interesse nessa pesquisa é especificamente em dados abertos do Arquivo Nacional.

3 METODOLOGIA DE PESQUISA

A pesquisa possui uma natureza aplicada, para gerar conhecimentos voltados para a aplicação prática e a solução de problemas específicos, envolvendo interesses

⁴ Abordagem parte de decisões centralizadas *top-down* (de cima para baixo).

⁵ Abordagem colaborativa e participativa, *bottom-up* (de baixo para cima).

locais (Prodonav; Freitas, 2013, p. 51). Quanto ao objetivo, a pesquisa é exploratória (Gil, 2002, p. 41), e a abordagem adotada é qualitativa (Prodonav; Freitas, 2013, p. 70).

Os procedimentos técnicos incluem a pesquisa bibliográfica e a pesquisa documental, pautada no estudo de caso. A pesquisa bibliográfica é elaborada a partir de material já publicado, como livros, revistas, publicações em periódicos e artigos científicos (Prodonav e Freitas, 2013, p. 54). A pesquisa documental, por sua vez, utiliza documentos, sejam escritos ou não, constituindo fontes primárias que podem ser coletadas quando o fato ou fenômeno ocorre ou posteriormente (Lakatos; Marconi, 2003, p. 174).

Para a coleta de dados, foram utilizados: o Google Acadêmico, a base de dados especializada *Scopus* e a *Brapi*, além de bibliotecas digitais como *SciELO*. Além disso, utilizamos uma planilha de dados abertos do Arquivo Nacional, através do Registro de entrada de documentos no Arquivo Nacional.

3.1 Local de Pesquisa e Dados Selecionados

O local da pesquisa, o Arquivo Nacional, é uma instituição criada em 1838 e atua como órgão central do Sistema de Gestão de Documentos e Arquivos (SIGA) da administração pública federal, integrando a estrutura do Ministério da Gestão e da Inovação em Serviços Públicos (Brasil, 2022).

Os dados utilizados foram obtidos através da base “Registro de entrada de acervo”, disponível para download em diversos formatos (.xlsx, .pdf, .ods e .json).⁶ Segue no Quadro 1 um recorte de 5 linhas da planilha, para a representação visual dos dados obtidos.

Quadro 1 - Planilha base registro de entrada de acervo (recorte de 5 linhas)

anoEntrada	Numero Registro	codigoReferencia	NomeFund	DataEntrada	FormaEntrada	Procedencia	Processo
2024	1	BR DFANBSB DW	MINISTÉRIO DA SAÚDE	01/01/2024	recolhim ento	Ministério da Saúde	08227.3884/2022-25
2024	2	BR AN BSB DW	MINISTÉRIO DA SAÚDE	01/01/2024	recolhim ento	Ministério da Saúde	08227.3884/2022-25
2024	3	BR DFANBSB 4V	MINISTÉRIO DA JUSTIÇA	01/01/2024	recolhim ento	Ministério da Justiça	08227.1716/2022-03

⁶ Planilha eletrônica da Bases de Dados Abertas PDA 2022-2024. Encontra-se na Página do Arquivo Nacional, na seção “Base registro de entrada de acervo – Março/2024”. Disponível em: <https://www.gov.br/arquivonacional/pt-br/acesso-a-informacao/dados-abertos-2/bases-de-dados-abertos-pda-2022-2024/REGENTES.xlsx>. Acesso em: 19 set. 2024.

2023	2	BR DFANBSB ARX	OBJETO VOADOR NÃO IDENTIFICAD O	31/03/2023	sem especific ação	Centro de Documentaç ão da Aeronáutica	00322.324/2 009-51
2023	4	BR RJANRIO L7	ALEXANDRE NIEMEYER	01/01/2023	comodat o	Niemeyer, Alexandre	00320.179/2 006-86

Fonte: Brasil (2024).

A planilha original possui os metadados organizados em 8 colunas (descritas no Quadro 2) e os dados distribuídos em 4435 linhas, antes de fazermos a limpeza de dados.

Quadro 2 - Dicionário de dados

METADADO	DESCRIÇÃO
anoEntrada	Ano referente ao recolhimento do fundo para o Arquivo Nacional (AN);
NumeroRegistro	Número sequencial de registro da entrada no acervo, do início do ano em janeiro até o fechamento em dezembro;
codigoReferencia	O código do país (BR), o código da entidade custodiadora e o código específico da unidade de descrição (Nobrade, 2006, p. 21).
NomeFundo	Nome do conjunto de documentos produzidos ou reunidos pelas instituições ao longo de suas atividades
DataEntrada	Data do recolhimento do fundo ao Arquivo
FormaEntrada	Critério utilizado para definir a forma de agregação do fundo ao Arquivo
Procedencia	Instituição que cedeu os documentos ao Arquivo
Processo	Dossiê contendo documentos que foram enviados ao AN

Fonte: Elaboração Própria.

3.2 Ferramentas e Limpeza dos Dados

As ferramentas utilizadas foram o *Google Sheets* e o *Power BI*. O *Google Sheets*, ou Planilhas Google, é um aplicativo de planilhas online gratuito que permite criar, editar e formatar arquivos, além de colaborar com outras pessoas em tempo real. O *Power BI* é uma ferramenta de análise de dados e negócios desenvolvida pela *Microsoft*. É composta por uma coleção de aplicativos, serviços de software e conectores que transformam fontes de dados em informações visualmente atraentes e interativas (Google, 2024).

No *Google Sheets*, os dados foram utilizados na íntegra. No entanto, antes de submeter os dados ao Power BI, foi necessário prepará-los. Na etapa de limpeza dos dados, foi utilizada a ferramenta *OpenRefine*⁷, da seguinte forma:

⁷ Disponível em <https://openrefine.org/>. Acesso em 14 abr. 2025.

- a) A planilha apresentava-se com células consideradas nulas pelo *Google Sheets*, essas foram retiradas no tratamento dos dados usando *Open Refine*, quando estavam localizadas nas colunas identificadas como processo, significando a não existência de documento;
- b) a planilha original continha 4.435 linhas. Após a limpeza dos dados, foram excluídas 2.638 linhas, restando 1797 linhas no total a serem utilizadas.

4 RESULTADOS E DISCUSSÃO

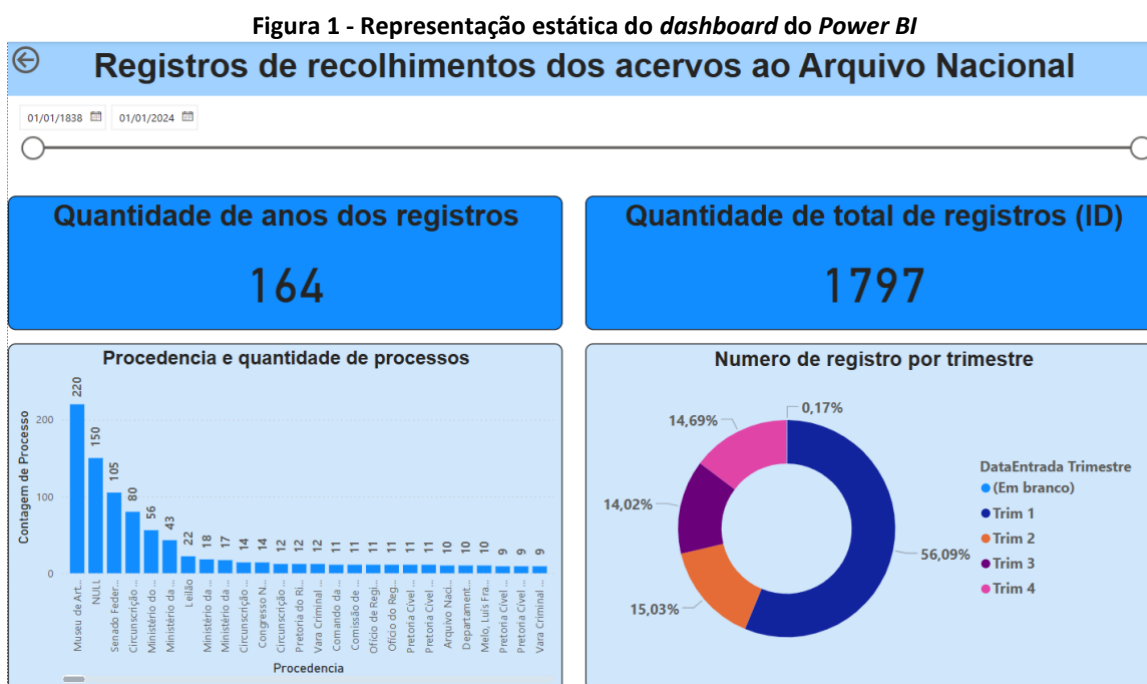
Desenvolvemos um diagnóstico inicial com os recursos do *Google Sheets*. O Gráfico 1 demonstra abaixo um recorte desse diagnóstico, representando os 10 anos que mais foram recolhidos acervos.



Fonte: Elaboração própria.

Notamos que os gráficos criados no *Google Sheets* possuem uma limitação da representação das categorias de uma planilha com grande extensão, por exemplo, a extensão de anos que a planilha espelha torna difícil a visualização em um único gráfico, por isso o Gráfico 1 foi restrito a apenas 10 anos. Logo, notamos a necessidade de uma outra ferramenta que sintetize essa representatividade. A ferramenta escolhida foi o *Power BI* da *Microsoft*. O *dashboard* interativo apresenta gráficos sobrepostos que permitem uma análise mais profunda das informações contidas nos metadados da planilha analisada.

Na Figura 1 mostramos uma versão estática do *dashboard*. A versão interativa dos “Registros de recolhimentos dos acervos ao Arquivo Nacional”⁸ está disponível online.⁹



Fonte: Elaboração própria.

A análise dos dados do *dashboard* da Figura 1 é apresentada a seguir:

- Inicialmente, consideramos que o metadado **anoEntrada**, que representa a **quantidade de anos de registros**, foi útil para entender quantos anos distintos o Arquivo Nacional recolheu acervos, totalizando 164 anos de registros. É importante destacar que essa quantidade não representa um simples somatório dos anos, mas sim uma contagem dos anos distintos entre o primeiro e o último recolhimento;
- o cartão ao lado direito exibe a **quantidade total de registros**, um dado que também é interativo. Após o processo de limpeza de dados com o *OpenRefine*, utilizamos os dados que exibe o número de registros "limpos", ou seja, o número das linhas válidas, correspondendo à quantidade de processos avaliados sem nulidades, totalizando 1797 unidades (**processos**);

⁸ A versão interativa pode ser acessada no link <https://encurtador.com.br/Ffrwm>.

⁹ Caso a visualização interativa não abra online, é possível baixar o arquivo “Recolhimento de acervos ao Arquivo Nacional.pbix” para sua máquina e executá-lo em uma versão local de *Power BI*.

- c) o gráfico **procedência e quantidade de processo**, representa quem cedeu o fundo e a quantidade de processos que cada órgão ou instituição enviou ao Arquivo Nacional. Esse gráfico foi desenvolvido com base nos metadados das colunas **procedência** e **processo**;
- d) o gráfico que apresenta o **número de registros por trimestre**, utiliza os metadados colunas **DataEntrada** e **NúmeroRegistro**. Ele apresenta as entradas trimestrais, expressas em percentuais, evidenciando que a maior parte dos documentos é recolhida no primeiro trimestre de cada ano. Essa informação pode auxiliar no planejamento e no desenvolvimento de estratégias administrativas para atividades futuras de recolhimento;
- e) por fim, também elaboramos uma **linha do tempo** interativa com os demais gráficos, estruturada através do metadado **DataEntrada**. Ela permite análises tanto de curtos quanto longos períodos, oferecendo uma base para traçar diretrizes no contexto da Governança de Dados.

Um exemplo de leitura do *dashboard* para orientar políticas públicas é que no gráfico que apresenta o número de registros por trimestre observa-se que no primeiro trimestre são recebidos mais da metade dos processos, sendo assim, o Arquivo Nacional pode se planejar para ter mais pessoal ativo neste período para receber e processar a documentação recebida. Como na barra superior é possível ajustar a data dos dados apresentados nos gráficos abaixo, é possível fazer recortes mais atuais para verificar se essa tendência de receber mais processos no primeiro trimestre se mantém nos anos mais recentes e se ainda justifica-se o planejamento exemplificado acima.

5 CONSIDERAÇÕES FINAIS

Apresentamos uma abordagem referente a Governança de Dados no contexto de Cidades Inteligentes, com cerne na reutilização de dados culturais e a aplicação de tecnologias para a difusão do conhecimento. Inicialmente, foi apresentado o papel fundamental da Governança de Dados como uma estrutura que promove a integridade, a disponibilidade e a conformidade dos dados, essenciais para a eficiência da gestão

urbana em Cidades Inteligentes. Através dessa governança, é possível garantir que os dados sejam tratados de maneira responsável, segura, sustentável e social.

Em seguida, exploramos como as Cidades Inteligentes podem se beneficiar do uso de dados provenientes dos patrimônios culturais. Estes dados, que são atualmente subutilizados, possuem grande potencial para fomentar a disseminação de informações sociais, educacionais e turísticas. Ao conectar o patrimônio cultural ao universo dos dados, criam-se oportunidades para que a sociedade tenha um acesso mais amplo a informações históricas e memoriais, contribuindo para a valorização e preservação da memória coletiva e individual.

Demonstrou de maneira prática como esses dados podem ser reaproveitados para disseminar conhecimento de forma eficiente e interativa, utilizando ferramentas como *dashboards* interativos. A visualização facilitada e acessível de informações permite uma análise mais profunda dos dados, incentivando uma tomada de decisões mais informada e colaborativa. Ao integrar esses dados em plataformas dinâmicas como o *Power BI*, torna-se possível traçar diretrizes mais eficazes para o planejamento e a gestão de projetos culturais em Cidades Inteligentes, ampliando o impacto desses dados no cotidiano urbano.

Verificamos que a Governança de Dados, aliada ao uso de dados culturais e à tecnologia, oferece um caminho permitindo que a cultura e a história não sejam apenas preservadas, mas também continuamente exploradas e reutilizadas em prol da sociedade.

No contexto das Cidades Inteligentes, inicialmente foram procurados dados culturais abertos nos municípios do Espírito Santo, buscando se aproximar da realidade local. Devido à dificuldade de localizar tais dados, foi utilizada uma planilha do Arquivo Nacional, que também proporcionou insights valiosos sobre a reutilização de dados culturais. Esse desafio pode integrar os projetos de Cidades Inteligentes, evidenciando a necessidade de um enfoque mais direto nas práticas culturais, nas manifestações, nos equipamentos culturais e na preservação de bens patrimoniais históricos.

Sugerimos que municípios e estados adotem abordagens semelhantes, utilizando dados abertos culturais, que facilitam a visualização e a análise de informações,

contribuindo para o planejamento e a gestão da cultura no contexto das Cidades Inteligentes.

Com base nessa experiência, sugerimos como insights para pesquisas futuras, a atenção para a qualidade dos dados (*data quality*), ou seja, utilizar ferramentas de gestão de qualidade de dados para identificar e corrigir erros e falhas no início da produção de dados, para evitar falhas conforme verificadas na planilha utilizada.

REFERÊNCIAS

BARATA, André M. **Governança de Dados em organizações brasileiras**: uma avaliação comparativa entre os benefícios previstos na literatura e os obtidos pelas organizações. 2015. Dissertação (Mestrado em Ciências) – Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2015. Disponível em <https://www.teses.usp.br/teses/disponiveis/100/100131/tde-28072015-215618/pt-br.php>. Acesso em: 20 maio 2025.

BENITES, Ana J. **Análise das Cidades Inteligentes sob a perspectiva da sustentabilidade**: o caso do centro de operações do Rio de Janeiro. 2016. 224 f. Dissertação (Mestrado em Política Científica e Tecnológica) – Instituto de Geociências, Universidade Estadual de Campinas, Campinas, 2016. Disponível em: <https://repositorio.unicamp.br/acervo/detalhe/978589>. Acesso em: 20 maio 2025.

BOUSKELA, M. *et al.* **Caminho para as Smart Cities**: da gestão tradicional para a Cidade Inteligente. Brasília: Banco Interamericano de Desenvolvimento (BID), 2016. Disponível em: <https://publications.iadb.org/publications/portuguese/document/Caminho-para-as-smart-cities-Da-gest%C3%A3o-tradicional-para-a-cidade-inteligente.pdf>. Acesso em: 20 jul. 2023.

BRASIL. Arquivo Nacional. **Bases de Dados Abertas PDA 2022-2024**. Brasília: Arquivo Nacional, mar. 2024. Planilha eletrônica. Disponível em: <https://www.gov.br/arquivonacional/pt-br/aceso-a-informacao/dados-abertos-2/bases-de-dados-abertos-pda-2022-2024/REGENTES.xlsx>. Acesso em: 19 set. 2024.

BRASIL. Decreto n. 8.777, de 11 de maio de 2016. Institui a Política de Dados Abertos do Poder Executivo Federal. **Diário Oficial da União**: seção 1, Brasília, DF, n. 90, p. 21, 12 maio 2016. Disponível em: http://www.planalto.gov.br/ccivil_03/ato2015-2018/2016/decreto/d8777.htm. Acesso em: 14 abr. 2025.

BRASIL. Lei n. 14.129, de 29 de março de 2021. Dispõe sobre princípios, regras e instrumentos para o Governo Digital e para o aumento da eficiência pública e dá outras providências. **Diário Oficial da União**: seção 1, Brasília, DF, n. 60, p. 3, 30 mar. 2021. Disponível em: <https://www.in.gov.br/en/web/dou/-/lei-n-14.129-de-29-de-marco-de->

[2021-311282132](#). Acesso em: 25 jul. 2022.

BRASIL. Serviços e informações do Brasil. **Arquivo Nacional (NA)**. Brasília, 13 maio 2022. Disponível em: <https://www.gov.br/pt-br/orgaos/arquivo-nacional>. Acesso em: 17 set. 2024.

CASTRO, J. M.; BARACHO, R. M. A. O patrimônio cultural nas cidades inteligentes. **Em Questão**, Porto Alegre, v. 26, n. 3, p. 298-326, set./dez. 2020.

GIL, Antônio Carlos. **Como elaborar projetos de pesquisa**. São Paulo: Atlas, 2002.

LAKATOS, Eva Maria; MARCONI, Marina de Andrade. **Fundamentos de metodologia científica**. São Paulo: Atlas, 2003.

MENDES, Teresa C. M. **Smart Cities**: solução para as cidades ou aprofundamento das desigualdades sociais? Rio de Janeiro: Observatório das Metrôpoles (INTC), 2020. (Texto para Discussão, n. 011/2020). Disponível em: https://www.observatoriodasmetrolopes.net.br/wp-content/uploads/2020/01/TD-011-2020_Teresa-Mendes_Final.pdf. Acesso em: 3 set. 2024.

PRODANOV, Cleber Cristiano; FREITAS, Ernani Cesar de. **Metodologia do trabalho científico**: métodos e técnicas da pesquisa e do trabalho. Novo Hamburgo: Feevale, 2013.

RÊGO, Bergson Lopes. **Gestão e Governança de Dados**: promovendo dados como ativo de valor nas empresas. Rio de Janeiro: Brasport, 2013. *E-book*. Disponível em: <https://plataforma.bvirtual.com.br>. Acesso em: 3 set. 2024.

THE DATA GOVERNANCE INSTITUTE (DGI). DGI Data Governance Framework. **Framework component #1**: mission and value. 2024. Disponível em: <https://datagovernance.com/the-dgi-data-governance-framework/>. Acesso em: 1 set. 2024.



[4.0 Internacional](#)

Esta licença permite compartilhamento, remixe, adaptação e criação a partir do trabalho, mesmo para fins comerciais, desde que sejam atribuídos créditos ao(s) autor(es). Conteúdos elaborados por terceiros, citados e referenciados nesta obra não são cobertos pela licença.

Como citar este trabalho:

SILVA, Carlos Luiz; FERRARI, Luciana Itida. Reuso de dados abertos culturais na Governança de Dados em Cidades Inteligentes: estudo com dados do Arquivo Nacional. In: WORKSHOP DE INFORMAÇÃO DADOS E TECNOLOGIA, 8., 2025, Marília, SP. **Anais [...]**. Marília, SP: Universidade de Marília, 2025. DOI: <https://doi.org/10.22477/viii.widat.274>.



WIDaT 2025

VIII WORKSHOP DE INFORMAÇÃO,
DADOS E TECNOLOGIA

MARÍLIA • SP

Linked Art Data Model: perfil de aplicação para descrever recursos artístico-culturais

Linked Art Data Model: application profile for describing artistic-cultural resources

Laís Barbudo Carrasco¹



<http://lattes.cnpq.br/7793712097080187>



<https://orcid.org/0000-0003-2938-9390>

Silvana Aparecida Borsetti Gregorio Vidotti²



<http://lattes.cnpq.br/7390573927636069>



<https://orcid.org/0000-0002-4216-0374>

Resumo

A representação de objetos e conteúdos artístico-culturais em museus de arte enfrenta desafios devido à diversidade de dados e à complexidade intrínseca dos conteúdos e obras de arte, impactando o acesso, a preservação e a pesquisa. Nesse contexto, o modelo de dados Linked Art, baseado no CIDOC CRM, visa padronizar e interconectar dados culturais, promovendo interoperabilidade, reuso de dados e enriquecimento da experiência do usuário. O artigo busca fornecer uma visão abrangente do Linked Art e suas potencialidades na representação digital de objetos e conteúdos de arte em ambientes digitais de museus, destacando oportunidades e desafios associados. Utilizando uma metodologia qualitativa e exploratória, a pesquisa realiza uma revisão bibliográfica e analisa o conteúdo técnico do Linked Art para compreender seus componentes, escopo e aplicação, revelando oportunidades e desafios associados ao modelo. Mesmo com os desafios, como a complexidade dos metadados, integração de sistemas, necessidade de capacitação do pessoal e gestão de mudanças organizacionais, a implementação do Linked Art traz benefícios como padronização para interoperabilidade, (re)uso e enriquecimento de dados, ampliação e descoberta de conteúdos e aprimoramento da experiência do usuário. Em conclusão, ressalta-se o potencial transformador do modelo Linked Art na representação e contextualização das obras de arte, enfatizando a necessidade contínua de pesquisas colaborativas para maximizar seu impacto, eficácia e sustentabilidade. Recomendam-se estudos futuros sobre melhores práticas de implementação e avaliação de longo prazo do Linked Art nas instituições culturais, e na experiência do público e de pesquisadores.

Palavras-chave: Linked Art; CIDOC CRM; metadados; museus de arte; representação digital.

Abstract

The digital representation of objects and artistic-cultural content in art museums faces challenges due to the diversity of data and the intrinsic complexity of the content and artworks, impacting access, preservation, and research. In this context, the Linked Art data model, based on CIDOC CRM, aims to

¹ Doutora em Ciência da Informação, Universidade Estadual Paulista Júlio de Mesquita Filho (Unesp), Marília, São Paulo, Brasil. Professora do Programa de Pós-graduação em Ciência da Informação e Arquivologia, Mid Sweden University (MIUN), Sundsvall, Västernorrland, Suécia. lais.barbudocarrasco@miun.se.

² Doutora em Educação, Universidade Estadual Paulista Júlio de Mesquita Filho (Unesp), Marília, São Paulo, Brasil. Professora do Departamento e do Programa de Pós-graduação em Ciência da Informação, Universidade Estadual Paulista Júlio de Mesquita Filho (Unesp), Marília, São Paulo, Brasil. Coordenadora de Tecnologias Aplicadas (COTEA), Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), Brasília, Distrito Federal, Brasil. silvana.vidotti@unesp.br.

standardize and interconnect cultural data, promoting interoperability, data reuse, and the enhancement of user experience. This article seeks to provide a comprehensive overview of Linked Art and its potential in the digital representation of art objects and content within digital museum environments, highlighting associated opportunities and challenges. Using a qualitative and exploratory methodology, the research conducts a literature review and analyzes the technical content of Linked Art to understand its components, scope, and application, revealing opportunities and challenges associated with the model. Despite challenges such as metadata complexity, system integration, the need for staff training, and organizational change management, the implementation of Linked Art offers benefits such as standardization for interoperability, data (re)use and enrichment, content expansion and discovery, and improved user experience. In conclusion, the transformative potential of Linked Art in the representation and contextualization of artworks is emphasized, underscoring the ongoing need for collaborative research to maximize its impact, effectiveness, and sustainability. Future studies are recommended on best practices for implementation and long-term evaluation of Linked Art in cultural institutions, as well as its impact on audience and researcher experiences.

Keywords: Linked Art; CIDOC CRM; metadata; art museums; digital representation.

1 INTRODUÇÃO

A representação e documentação de objetos e conteúdos artístico-culturais em ambientes digitais têm se tornado um desafio para museus de arte e outras instituições culturais. A diversidade de formatos de dados, padrões de metadados e sistemas de informação dificulta a integração e a representação digital dos acervos. Além disso, as especificidades estéticas, históricas e culturais das obras de arte aumentam a complexidade do processo. A representação digital é fundamental para ampliar o acesso ao patrimônio cultural, contribuindo para seu acesso, preservação, educação e pesquisa. Esse desafio é particularmente relevante para museus que buscam oferecer experiências digitais mais acessíveis, imersivas e integradas.

Nesse contexto, o Linked Art Data Model, um perfil de aplicação de metadados baseado no CIDOC Conceptual Reference Model (CIDOC CRM), desenvolvido pelo Comitê Internacional de Documentação (CIDOC) do Conselho Internacional de Museus (ICOM), surge como uma solução promissora (Raemy; Sanderson, 2024; Geddes, 2019; Fink, 2018; Raemy, 2023; Newbury, 2018). A iniciativa visa criar uma estrutura técnica e conceitual que possibilite a interconexão de dados sobre obras de arte, artistas, exposições, instituições culturais e eventos, promovendo a interoperabilidade e o compartilhamento de informações.

Um perfil de aplicação de metadados define diretrizes e especificações para estruturar metadados em um domínio específico, garantindo consistência e interoperabilidade entre sistemas e repositórios. Esses perfis adaptam padrões existentes

para atender às necessidades específicas do domínio, facilitando a troca e gestão de metadados e contribuindo para uma recuperação de informações mais eficiente (Gerontakos; Riesenber, 2021).

Linked Art é uma comunidade que trabalha em conjunto para criar um Modelo compartilhado baseado em Dados Abertos e Utilizáveis Interligados (Linked Open Usable Data), com o objetivo de descrever o patrimônio cultural. O foco principal está nas obras de arte, mas inclui também arquivos e materiais bibliográficos. (Linked Art, s.d., s.p, tradução nossa)³

O Linked Art Data Model emprega princípios da Web Semântica e Linked Data para vincular e enriquecer dados culturais, tornando-os mais acessíveis e utilizáveis para aplicações como sistemas de gestão de museus, aplicativos de turismo cultural e pesquisas acadêmicas. Essa abordagem permite uma navegação e pesquisa mais eficiente e integrada das informações sobre arte e cultura.

A comunidade do Linked Art trabalha colaborativamente para criar um modelo baseado em Dados Abertos Interligados, focado na descrição do patrimônio cultural (Linked.Art, s.d.). Esse modelo, fundamentado no CIDOC CRM, tem um escopo mais restrito, concentrando-se em museus de arte em vez de abordar todo o gerenciamento do patrimônio cultural. Sua adoção impulsiona a implementação do CIDOC CRM e gera novos requisitos para o CIDOC CRM SIG (Special Interest Group), promovendo um intercâmbio dinâmico de conhecimento (Linked Art - CIDOC CRM Working Group, s.d.).

O Linked Art Data Model também promove a interoperabilidade entre sistemas, facilitando o compartilhamento de dados culturais de forma padronizada e acessível. Isso melhora a experiência do público ao permitir acesso integrado a informações sobre o patrimônio cultural (Raemy; Sanderson, 2024; Geddes, 2019; Fink, 2018; Raemy, 2023; Newbury, 2018). Dessa forma, a iniciativa democratiza o acesso ao conhecimento cultural e amplia o impacto das coleções de arte globalmente.

O Linked Art Data Model destaca-se pelo seu potencial em estruturar a representação digital de obras de arte e objetos culturais, permitindo a interconexão de diferentes tipos de dados e proporcionando uma visão contextualizada e integrada dos acervos. Sua comunidade inclui renomadas instituições acadêmicas, museológicas e

³ Linked Art is a community working together to create a shared model based on Linked Open Usable Data to describe cultural heritage, with a focus on artwork but including also archives and bibliographic material (Linked Art, s.d.).

tecnológicas, contribuindo para o desenvolvimento de padrões interoperáveis no campo da informação cultural. Entre seus membros destacam-se o J. Paul Getty Trust, na padronização de metadados; o Smithsonian Institution, na agregação e disseminação de coleções digitais; e a Europeia, no acesso e reutilização de dados culturais. A rede também conta com universidades como Harvard e Oxford e museus como o Louvre e o Metropolitan Museum of Art (Linked.Art Community, s.d.).

A complexidade da representação digital de objetos artísticos influencia a formação de narrativas culturais, a criação de experiências digitais e a preservação do conhecimento. Embora estudos tenham explorado o CIDOC CRM em contextos de patrimônio cultural, ainda há uma lacuna na literatura brasileira sobre os desafios e oportunidades do Linked Art Data Model e sua implementação em instituições brasileiras.

Neste contexto, este estudo tem como objetivo analisar o Linked Art Data Model e suas possíveis contribuições para a representação digital de acervos artístico-culturais, considerando desafios e oportunidades. A pesquisa explora as implicações dessa abordagem para instituições culturais e promove um diálogo com curadores, gestores, pesquisadores, museólogos, bibliotecários, arquivistas e desenvolvedores de sistemas de informação cultural.

2 PROCEDIMENTOS METODOLÓGICOS

Para atingir os objetivos da pesquisa, adotou-se uma abordagem qualitativa e exploratória, que incluiu revisão bibliográfica, análise da literatura e estudo do conteúdo técnico do Linked Art. Essa metodologia permitiu analisar e compreender os princípios, diretrizes e potencialidades do modelo, estabelecendo uma base conceitual para avaliar o Linked Art Data Model e sua relevância no contexto da catalogação e representação do patrimônio cultural em museus de arte.

3 RESULTADOS

O Linked Art Data Model surgiu como resposta à crescente necessidade de integrar dados culturais em ambientes digitais, especialmente em museus, galerias e instituições do setor. Trata-se de uma iniciativa colaborativa da comunidade profissional

de museus e patrimônio cultural, que desenvolve um modelo compartilhado baseado em Dados Abertos Interligados (Linked Open Data – LOD) para descrever o patrimônio cultural, com foco particular em obras de arte (Linked.Art, s.d.; Linked.Art Community, s.d.).

Linked Art é uma comunidade que trabalha em conjunto para criar um modelo compartilhado baseado em Dados Abertos Interligados para descrever o patrimônio cultural, com um foco particular em arte. Linked Art é uma maneira utilizável e baseada em padrões para descrever o patrimônio cultural. Ou, de forma mais detalhada, Linked Art é uma comunidade de profissionais de museus e patrimônio cultural que colaboram para definir um perfil de aplicação de metadados (o modelo) para descrever o patrimônio cultural, e os meios técnicos para interagir convenientemente com ele (a API). (Linked.Art, s.d., s.p, tradução nossa)⁴

A proposta do Linked Art é fornecer um perfil de aplicação de metadados e uma interface de programação de aplicações (API) que permitam a descrição estruturada de recursos culturais. O modelo adota uma abordagem centrada em eventos, fundamentada no CIDOC Conceptual Reference Model (CIDOC CRM), que facilita a representação de dados temporais e contextuais (Raemy, 2023; Raemy; Sanderson, 2024).

Essa abordagem é complementada pelo uso de vocabulários controlados do Getty, incluindo o *Art & Architecture Thesaurus* (AAT), o *Getty Thesaurus of Geographic Names* (TGN) e a *Union List of Artist Names* (ULAN) (Raemy, 2023). O Linked Art fornece um perfil de aplicação de metadados e uma API (*Application Programming Interface*) para descrever o patrimônio cultural, utilizando uma ontologia centrada em eventos para modelar dados temporais de maneira eficaz (Raemy, 2023). Priorizando usabilidade e consistência, a iniciativa busca atender à maioria dos casos de uso com complexidade mínima, considerando as necessidades dos desenvolvedores que trabalham com dados de patrimônio cultural (Raemy; Sanderson, 2024).

Para simplificar a modelagem e o consumo de dados, o Linked Art adota o JSON-LD (*JavaScript Object Notation for Linked Data*) como sintaxe, favorecendo a facilidade de uso em detrimento de algumas funcionalidades do RDF (*Resource Description Framework*)

⁴ Linked Art is a community working together to create a shared Model based on Linked Open Data to describe cultural heritage with a particular focus on art. Linked Art is a usable, standards-based way to describe cultural heritage. Or, in more detail, Linked Art is a community of museum and cultural heritage professionals collaborating to define a metadata application profile (the model) for describing cultural heritage, and the technical means for conveniently interacting with it (the API). (Linked.Art, s.d.).

(Newbury, 2018). A implementação envolve o mapeamento de classes e relações do CIDOC CRM para propriedades JSON e a incorporação de elementos de outros vocabulários RDF, garantindo interoperabilidade e consistência (Newbury, 2018). Assim, ao estabelecer princípios de design baseados em padrões comuns, o Linked Art beneficia editores e consumidores de dados, facilitando a extração e reutilização de informações (Newbury, 2018).

O Linked Art se destaca por seu modelo conceitual flexível e modular, que permite a representação de uma ampla gama de elementos do patrimônio cultural, como objetos físicos e digitais, pessoas, locais, exposições, atividades museológicas, ações de conservação, proveniência, documentos textuais e coleções. Essa estrutura, baseada em princípios semânticos, facilita a integração entre instituições e o compartilhamento eficiente e reutilizável de dados (Raemy; Sanderson, 2024; Geddes, 2019; Fink, 2018; Newbury, 2018). Fundamentado no CIDOC CRM, o modelo utiliza um subconjunto simplificado para garantir consistência e adota os Vocabulários Getty como referência para a terminologia específica do domínio (Linked Art, s.d.).

O Quadro 1 apresenta as classes principais no contexto do Linked Art, cada uma com uma breve descrição.

Quadro 1 – Classes principais do Linked Art

Classe	Descrição
Human Made Object	Objetos físicos tangíveis, como uma pintura.
Digital Object	Objetos que existem apenas digitalmente, como uma página web.
Person	Pessoa humana, por exemplo, Rembrandt.
Group	Grupo de pessoas com ação coletiva, como um departamento de museu.
Place	Local com coordenadas geográficas, como Paris.
Visual Item	Conteúdo visual ou imagem, como a representação da “Mona Lisa”.
Linguistic Object	Conteúdo textual, como o texto de <i>O Senhor dos Anéis</i> .
Propositional Object	Obra abstrata não visual ou linguística, como uma exposição conceitual.
Type	Categoria ou conceito, como “centímetros” ou “escultura”.

Classe	Descrição
Set	Conjunto ou coleção de entidades, como uma coleção de museu.
Activity	Ação realizada por pessoas ou grupos, como uma exposição ou transferência.

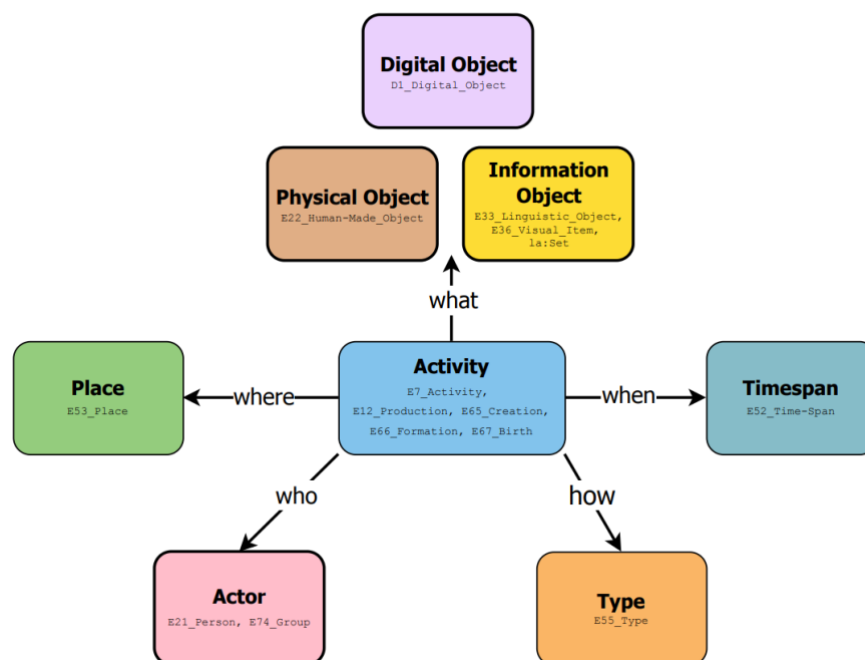
Fonte: Adaptado de Linked Art (s.d.)

A classe “Human Made Object” refere-se a objetos físicos tangíveis, como pinturas, enquanto “Digital Object” engloba entidades puramente digitais, como páginas da web. “Person” representa indivíduos humanos, como o artista Rembrandt, enquanto “Group” descreve conjuntos de pessoas capazes de ação coletiva, como departamentos de museus. “Place” identifica locais geográficos, como a cidade de Paris, com a possibilidade de coordenadas. “Visual Item” refere-se a conteúdos visuais ou imagens conceituais associadas a objetos físicos ou digitais, como a famosa obra “Mona Lisa”. Por sua vez, “Linguistic Object” designa conteúdos textuais transportados por objetos físicos ou digitais, exemplificado pelo texto da obra O Senhor dos Anéis. A classe “Propositional Object” representa obras abstratas que não são nem linguísticas nem visuais, como conceitos de exposições ou peças de arte performática. “Type” refere-se a categorias ou conceitos, como unidades de medida ou tipos de esculturas, enquanto “Set” engloba coleções de outras entidades, como conjuntos de obras de arte em um museu. Por fim, “Activity” representa ações realizadas por pessoas ou grupos, como exposições ou transações de propriedade de objetos entre partes.

A Figura 1 representa o modelo conceitual do Linked Art, o qual utiliza as classes do CIDOC CRM e aborda cinco perguntas fundamentais sobre a proveniência. O diagrama mostra como essas classes e suas relações são organizadas para responder a cinco perguntas-chave relacionadas à proveniência de um objeto cultural: o que (what), onde (where), quem (who), como (how) e quando (when). Essas perguntas abrangem aspectos essenciais do histórico e do contexto de um objeto cultural, ajudando a compreender sua origem e trajetória ao longo do tempo.

Ao estruturar coleções de arte e patrimônio cultural como *Linked Open Data* (LOD) e utilizar RDF, o Linked Art proporciona uma representação organizada das entidades e suas relações. Essa abordagem permite descrições detalhadas de objetos culturais que abrangem metadados técnicos, históricos e contextuais.

Figura 1 – O modelo Linked Art



Fonte: Raemy e Sanderson (2024, p. 10).

Ao equilibrar flexibilidade, consistência e padrões abertos, o Linked Art Data Model contribui para um ecossistema de dados mais integrado no campo do patrimônio cultural. Sua estrutura permite o compartilhamento eficiente de informações entre instituições, aprimorando a gestão, a pesquisa e a difusão de acervos culturais em ambientes digitais. Sendo assim, o Linked Art representa um avanço significativo na padronização da descrição e interconexão de dados culturais. Com base em uma abordagem colaborativa e centrada em eventos, o modelo fortalece a interoperabilidade entre instituições, promove o engajamento com recursos culturais e apoia a criação de uma memória cultural mais acessível, conectada e duradoura.

Quadro 2 – Linked Art: oportunidades e desafios

Oportunidades	Desafios
Padronização e Interoperabilidade	Complexidade dos Metadados
Facilita o compartilhamento de dados entre sistemas e instituições. Amplia acesso e visibilidade das coleções.	Criação e manutenção de metadados estruturados são complexas. Necessidade de garantir qualidade e consistência.
Enriquecimento da Experiência do Usuário	Integração de Sistemas
Cria experiências digitais interativas e contextualizadas.	Desafios técnicos ao integrar com sistemas existentes.

Proporciona compreensão aprofundada das coleções.	Desafios operacionais na implementação de novas tecnologias.
Acesso Global	Aprendizado e Capacitação
Amplia o acesso a coleções, especialmente para públicos remotos. Facilita a exploração de obras em plataformas digitais.	Investimento em treinamento para uso de diretrizes e ferramentas do Linked Art.
Integração de Dados	Gestão de Mudanças
Promove enriquecimento, reuso e catalogação compartilhada de dados.	Requer mudanças na cultura organizacional e processos. Planejamento estratégico para adoção sustentável.

Fonte: Elaboração própria.

A adoção do Linked Art por museus de arte apresenta tanto oportunidades quanto desafios. A padronização da representação e interconexão de dados culturais fortalece a interoperabilidade entre sistemas, facilitando o compartilhamento de informações entre instituições. Isso contribui para uma experiência mais enriquecedora, permitindo uma compreensão aprofundada das obras em seus contextos históricos, culturais e artísticos. Além disso, a iniciativa favorece o reuso de dados e a catalogação compartilhada, otimizando a gestão das coleções e promovendo maior acesso.

Por outro lado, a implementação do Linked Art Data Model exige investimentos em recursos humanos, tecnológicos e financeiros. A criação e manutenção de metadados estruturados e padronizados podem ser complexas, especialmente devido à diversidade das coleções. A integração com sistemas preexistentes pode representar desafios técnicos e operacionais, demandando adaptações e atualizações constantes.

Outro aspecto relevante é a necessidade de capacitação profissional. Para garantir a qualidade e consistência dos dados, os profissionais dos museus devem ser treinados nas diretrizes e ferramentas associadas ao Linked Art. Além disso, a adoção desse modelo implica mudanças na cultura organizacional e nos processos institucionais, exigindo planejamento estratégico e gestão eficaz para assegurar uma transição estruturada e sustentável.

4 CONSIDERAÇÕES FINAIS

A representação e documentação de objetos e conteúdos artístico-culturais em ambientes digitais apresentam desafios significativos para instituições culturais, especialmente museus de arte. A variedade de formatos de dados, padrões de metadados e sistemas de informação dificulta a criação de acervos digitais integrados e padronizados. Nesse cenário, o Linked Art tem sido explorado como um modelo promissor para superar essas dificuldades.

Este estudo analisa o Linked Art Data Model, um perfil de aplicação de metadados baseado no CIDOC Conceptual Reference Model (CIDOC CRM), desenvolvido para facilitar a interconexão de dados no campo da arte e da cultura. A pesquisa aborda como essa abordagem estrutura a representação digital de objetos artísticos e seus contextos, promovendo a interoperabilidade e o compartilhamento de informações entre instituições.

A análise das oportunidades e desafios associados ao Linked Art destaca benefícios como a padronização de metadados, o aumento do acesso global, a integração de coleções e a melhoria da experiência do usuário. Contudo, também identifica obstáculos, como a complexidade técnica, a adaptação de sistemas existentes, a necessidade de capacitação profissional e os impactos na gestão institucional.

Além de ser um modelo técnico, o Linked Art Data Model representa uma rede colaborativa que envolve profissionais de museus, especialistas em tecnologia e instituições culturais, todos empenhados em ampliar a interoperabilidade e o acesso aos dados. A participação ativa nessa comunidade pode facilitar a implementação do modelo e garantir sua sustentabilidade a longo prazo.

Este estudo contribui para a compreensão do Linked Art ao explorar seu potencial e suas limitações, oferecendo insights para sua aplicação em instituições culturais. Pesquisas futuras podem se aprofundar nas melhores práticas de adoção e avaliar o impacto institucional e social desse modelo a longo prazo.

REFERÊNCIAS

CIDOC CRM. Special interest group. **CIDOC Conceptual Reference Model (CRM)**. Versão 7.1.3. fev. 2024. Disponível em: <https://site2024.cidoc-crm.org/Version/version-7.1.3>. Acesso em: 01 mar. 2025.

CIDOC CRM. Special interest group. **Linked Art**: a framework for linked open usable data in the visual arts. Disponível em: <https://cidoc.mini.icom.museum/working-groups/linked-art/>. Acesso em: 01 mar. 2025.

FINK, Eleanor E. Overview and recommendations for good practices. **American Art Collaborative (AAC) Linked Open Data (LOD) Initiative**, v. 35, 2018.

GEDDES, Margo. **Strategies to support wider adoption of linked open data in smaller museums**. Maryland: Johns Hopkins University 2019. Disponível em: <https://jscholarship.library.jhu.edu/server/api/core/bitstreams/dad33a38-00ad-452f-ab87-3df8ea67f27a/content>. Acesso em: 01 mar. 2025.

GERONTAKOS, Theodore; RIESENBERG, Benjamin. Metadata application profiles. **Lib Technol Rep**, v. 57, n. 6, p. 1-39, 2021.

LINKED ART COMMUNITY. **Linked art community**. Disponível em: <https://linked.art/community/>. Acesso em: 01 mar. 2025.

LINKED ART. **Linked art**: a framework for linked open usable data in the visual arts. Versão 1.0. fev. 2025. Disponível em: <https://linked.art/>. Acesso em: 01 mar. 2025.

NEWBURY, David. LOUD: linked open usable data and linked.art. *In*: CIDOC ANNUAL CONFERENCE, 2018, Heraklion, GR. Proceedings [...]. Heraklion, GR: [s. n.], 2018. p. 1-11. Disponível em: https://cidoc.mini.icom.museum/wp-content/uploads/sites/6/2021/03/CIDOC2018_paper_153.pdf. Acesso em: 01 mar. 2025.

RAEMY, Julien A.; SANDERSON, Robert. Analysis of the usability of automatically enriched cultural heritage data. *In*: MORAL-ANDRÉS, Fernando; MERINO-GÓMEZ, Elena; REVIRIEGO, Pedro. **Decoding cultural heritage**: a critical dissection and taxonomy of human creativity through digital tools. Cham: Springer Nature Switzerland, 2024. p. 69-93.

RAEMY, Julien Antoine. **Characterising the IIIF and linked art communities**. 2023. Disponível em: https://hal.science/hal-04162572v1/file/Characterising_IIIF_and_Linked_Art.pdf. Acesso em: 25 maio 2025.



[4.0 Internacional](#)

Esta licença permite compartilhamento, remixe, adaptação e criação a partir do trabalho, mesmo para fins comerciais, desde que sejam atribuídos créditos ao(s) autor(es). Conteúdos elaborados por terceiros, citados e referenciados nesta obra não são cobertos pela licença.

Como citar este trabalho:

CARRASCO, Laís Barbudo; VIDOTTI, Silvana Aparecida Borsetti Gregorio. *Linked Art Data Model: perfil de aplicação para descrever recursos artístico-culturais*. In: WORKSHOP DE INFORMAÇÃO, DADOS E TECNOLOGIA, 8., Marília, SP. **Anais [...]**. Marília, SP: Universidade de Marília. 2025. DOI: <https://doi.org/10.22477/viii.widat.255>.

Uma abordagem preditiva da evasão na educação a distância por meio de técnicas de mineração de dados educacionais

A predictive approach to dropout in distance education through educational data mining

Wesley Rodrigues Sepúlveda¹



<http://lattes.cnpq.br/4152945701822073>



<https://orcid.org/0009-0002-6132-3017>

Edilson Ferneda²



<http://lattes.cnpq.br/2531761427648020>



<https://orcid.org/0000-0003-4164-5828>

Fernando Antônio Hello³



<http://lattes.cnpq.br/2518246712240428>



<https://orcid.org/0000-0003-1421-386X>

Joyce Siqueira⁴



<http://lattes.cnpq.br/7006325340635761>



<https://orcid.org/0000-0003-0246-8443>

Djennifer Hipólito Portugal Amâncio Kneip⁵



<http://lattes.cnpq.br/2963176937865068>



<https://orcid.org/0009-0007-5564-6940>

Resumo

A Educação a Distância tem se consolidado como a principal escolha dos estudantes no Brasil, impulsionada por um crescimento expressivo na última década. No entanto, esse avanço trouxe o desafio da evasão estudantil, que compromete o desenvolvimento educacional e social, afeta investimentos públicos e privados e limita o

¹ Mestre, Universidade Católica de Brasília (UCB), Brasília, DF, Brasil. Docente, Faculdade de Tecnologia e Inovação Senac-DF, Brasília, DF, Brasil. wesleyrsepulveda@gmail.com.

² Doutor, LIRMM-Université Montpellier II, França. Docente, Universidade Católica de Brasília (UCB), Brasília, DF, Brasil. eferneda@p.ucb.br.

³ Doutor, Universidade de Campinas (Unicamp), Campinas, SP, Brasil. Empresa Brasileira de Pesquisa Agropecuária, Brasília, DF, Brasil. hello@unicamp.br.

⁴ Doutora, Universidade de Brasília (UnB), Brasília, DF, Brasil. Docente, Universidade Católica de Brasília (UCB), Brasília, DF, Brasil. joycitta@gmail.com.

⁵ Graduanda, Universidade Católica de Brasília (UCB), Brasília, DF, Brasil. Assistente de Tecnologias Educacionais, Universidade Católica de Brasília (UCB), Brasília, DF, Brasil. kneipdjennifer@outlook.com.

crescimento individual e coletivo dos estudantes. Nesse contexto, a Mineração de Dados Educacionais (*Educational Data Mining - EDM*) surge como uma estratégia eficaz para desenvolver modelos preditivos que auxiliem na redução da evasão em cursos de EAD. Este estudo analisa o comportamento de estudantes evadidos em cursos de graduação de uma universidade do Centro-Oeste, utilizando dados do Ambiente Virtual de Aprendizagem (AVA) e do Sistema de Gestão Acadêmica (SGA) para apoiar decisões preventivas. O processo de análise seguiu as metodologias Knowledge Discovery in Database (KDD) e CRISP-DM, estruturado em quatro etapas: revisão da literatura, análise dos dados do AVA e SGA, aplicação da ferramenta Weka com cinco algoritmos classificadores (Bayes Net, SMO, J48, Naive Bayes e Multilayer Perceptron) e a proposição de ações gerenciais para mitigar a evasão. Os resultados demonstram que a análise dos dados nas primeiras quatro semanas do curso permite prever a evasão com precisão significativa, oferecendo um recurso estratégico para a gestão acadêmica e possibilitando intervenções preventivas focadas na permanência estudantil.

Palavras-chave: educação a distância; *Knowledge Discovery in Database (KDD)*; mineração de dados; *Educational Data Mining (EDM)*; predição de evasão.

Abstract

Distance Education (EAD) has become the primary choice for students in Brazil, driven by significant growth over the past decade. However, this expansion has brought the challenge of student dropout, which undermines educational and social development, impacts public and private investments, and limits both individual and collective growth. In this context, Educational Data Mining (EDM) emerges as an effective strategy for developing predictive models that help reduce dropout rates in distance learning programs. This study analyzes the behavior of students who have dropped out of undergraduate courses at a university in the Midwest region, using data from the Virtual Learning Environment (AVA) and the Academic Management System (SGA) to support preventive decision-making. The analysis followed the Knowledge Discovery in Database (KDD) and CRISP-DM methodologies, structured into four stages: a literature review, data analysis from the AVA and SGA, application of the Weka tool with five classification algorithms (Bayes Net, SMO, J48, Naive Bayes, and Multilayer Perceptron), and the proposal of management actions to mitigate dropout. The results show that analyzing data within the first four weeks of the course enables the prediction of dropout with significant accuracy, providing a strategic resource for academic management and enabling preventive interventions focused on student retention.

Keywords: distance education; Knowledge Discovery in Database (KDD); data mining; Educational Data Mining (EDM); dropout prediction.

1 INTRODUÇÃO

Com o crescimento contínuo da Educação a Distância (EAD), diversas pesquisas têm sido conduzidas para compreender e reduzir a evasão nessa modalidade. Contextualizando o cenário, o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), com base nos resultados do Censo da Educação Superior 2021, destaca que o Ensino a Distância cresceu 474% em uma década, mas especificamente, entre os anos 2011 e 2021 (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2022). O Censo EAD.BR 2022 (Associação Brasileira de Educação à Distância, 2024) indicou que 73% dos alunos ingressantes no ano de 2022 optaram pela EAD. Assim, segundo o Censo EAD.BR 2022 “A sociedade optou pela EAD” (Associação Brasileira de Educação à Distância, 2022, p. 109).

A partir de números tão expressivos em relação ao ingresso de estudantes na modalidade, o Censo EAD.BR 2022 também apresenta a difícil realidade referente à alta evasão dos estudantes, cujo índice se concentra na faixa de até 40% (Associação Brasileira de Educação à Distância, 2024).

Os dados evidenciam que a evasão é um dos principais desafios enfrentados pelas Instituições de Ensino Superior (IES) no Brasil. O Censo EAD.BR 2022 (Associação Brasileira de Educação à Distância, 2024) destaca que o controle da evasão impacta tanto o desenvolvimento educacional quanto o social, afetando investimentos públicos e privados e limitando o crescimento individual e coletivo dos estudantes.

Ainda no Censo EAD.BR 2022 (Associação Brasileira de Educação à Distância, 2024), há fatores conhecidos que contribuem para a evasão no ensino superior, entre eles, as dificuldades financeiras, a falta de adaptação aos modelos pedagógicos e os desafios no uso de ferramentas tecnológicas. O documento ressalta que "O controle da evasão no ensino superior deve ser compreendido como um elemento crucial para o desenvolvimento educacional e social" (Associação Brasileira de Educação à Distância, 2024, p. 89). A permanência dos estudantes está diretamente relacionada à equidade e à inclusão (Associação Brasileira de Educação à Distância, 2024, p. 90).

Retomando a frase inicial deste texto: diversas pesquisas vêm sendo realizadas a fim de mitigar esse fenômeno. Neste contexto, trataremos especificamente de soluções computacionais capazes de auxiliar os gestores na tomada de decisões sobre o tema. Discutiremos uma linha de pesquisa que utiliza Técnicas de Extração de Conhecimento em Bases de Dados (do inglês, *Knowledge Discovery in Databases* – KDD) para extrair do Ambiente Virtual de Aprendizagem (AVA) e do Sistema de Gestão Acadêmica (SGA), conhecimento sobre o perfil e o comportamento dos estudantes evadidos.

Fayyad, Piatetsky-Shapiro e Smyth (1996) definem o KDD como um processo não trivial de extração de informações implícitas, previamente desconhecidas e potencialmente úteis a partir dos dados armazenados em um banco de dados. Esse processo envolve a execução de cinco etapas: a seleção, com a escolha dos dados relevantes para a análise; o pré-processamento, com a limpeza e preparação dos dados; a transformação, com a conversão dos dados para formatos adequados às técnicas de mineração; a mineração de dados, com a

aplicação de algoritmos para identificar padrões significativos; e a interpretação e avaliação, com a análise e validação dos padrões descobertos. Trata-se de um processo iterativo e interativo, onde todas as etapas estão conectadas e envolvem tarefas e decisões a serem realizadas pelos usuários.

Neste universo em que a Ciência da Computação se mostra como área dominante, é importante destacarmos a importante interdisciplinaridade com a Ciência da Informação (CI). Gomes, Pimenta e Schneider (2019) realizaram estudo sobre a mineração de dados na pesquisa em CI, no qual, relatam os inúmeros estudos de como os algoritmos têm impactado os aspectos sociais, econômicos e culturais na sociedade e de como o estudo crítico sobre a influência algorítmica e como as novas tecnologias de informação impactam a sociedade são uma das principais vocações do campo da CI.

Técnicas de mineração de dados são cada vez mais usadas em ciências sociais e humanidades. Gomes, Pimenta e Schneider (2019) destacam a classificação como um elemento comum na mineração de dados, cujo objetivo é identificar padrões que unem ou separam os dados analisados. Cada processo de mineração resulta em novos dados ou atributos para os conjuntos analisados, e as classificações emergentes, especialmente as complexas, necessitam de sistemas para organizar os resultados.

As pesquisas sobre evasão incluíram técnicas de KDD, utilizando informações geradas e mantidas nas bases de dados educacionais, a fim de identificar o comportamento dos estudantes com tendência a evadir, emitir alertas e/ou fornecer informações importantes para que medidas preventivas possam ser tomadas.

O *Education Data Mining* (EDM) é considerado um sinônimo de KDD quando aplicado a bases educacionais, embora o KDD tenha um escopo mais amplo. O EDM tem sido usado para identificar padrões de evasão e medir a motivação estudantil, auxiliando na tomada de decisões para reduzir a evasão. Suas atividades incluem análises preditivas e descritivas, mas nesta pesquisa, o foco é preditivo.

Dessa forma, considerando que: (i) a evasão de estudantes é apontada como o maior obstáculo enfrentado pelos cursos da modalidade EAD; (ii) partindo do pressuposto de que é possível identificar os estudantes com tendência a evadir de forma preditiva, por meio de técnicas do KDD; e (iii) a importância de gerar informações que forneçam insumo ao processo

de tomada de decisão, o objetivo da pesquisa é à construção de um modelo preditivo de evasão de estudantes de Educação a Distância (EAD), utilizando técnicas de KDD aplicadas a dados gerados por um Sistema de Gestão Acadêmica (SGA) e pelo Ambiente Virtual de Aprendizagem (AVA).

A compreensão deste comportamento, obtida por meio da interpretação dos dados dos alunos, fornecerá subsídios para a implementação de ações de mitigação da evasão, contribuindo para a manutenção de índices satisfatórios de permanência dos estudantes.

Assim, em posse de um modelo de predição de evasão, os gestores poderão receber, em tempo hábil, informações relevantes para acompanhar e agir preventivamente no combate a evasão, podendo colocar-se à disposição dos estudantes com alguma dificuldade.

Diante do cenário desafiador da evasão, este estudo reforça a relevância do uso de técnicas de Mineração de Dados Educacionais (EDM) como ferramenta estratégica para apoiar a tomada de decisões em instituições de ensino superior. Assim, os próximos passos desta pesquisa envolvem a apresentação detalhada da metodologia empregada e dos resultados obtidos com a aplicação dos algoritmos classificadores.

2 PROCEDIMENTOS METODOLÓGICOS

Trata-se de uma pesquisa de natureza aplicada, com abordagem mista, combinando métodos quantitativos e qualitativos, e com objetivos exploratórios. A investigação possui um meio experimental, utilizando a técnica de revisão bibliográfica e o método de pesquisa de estudo de caso.

O estudo de caso foi realizado em uma Instituição de Ensino Superior do Centro-Oeste, com foco nos alunos dos cursos de Tecnologia da Informação. A análise abrange dados de dois períodos letivos, totalizando 110 estudantes que utilizam o Ambiente Virtual de Aprendizagem (AVA) e o Sistema de Gestão Acadêmica (SGA). O estudo envolveu a seleção de dados e interações dos estudantes no primeiro semestre, comparando com o status desses mesmos alunos no segundo semestre. A partir dessa seleção, é possível identificar a classe de 'evadidos', que é o foco deste trabalho, e analisar os comportamentos dos estudantes por meio das técnicas de *Knowledge Discovery from Data* (KDD).

Para a investigação da base de dados na busca do comportamento do estudante propenso a evadir, foi utilizado a metodologia *Cross Industry Standard Process for Data Mining* (CRISP-DM) que possui um conjunto de seis fases (compreensão do negócio, compreensão dos dados, preparação de dados, modelagem, avaliação e implantação) que devem ser cumpridas para cobrir todo o ciclo de vida de um projeto de *Data Mining* (Hotz, 2024; Piatetsky, 2014). O CRISP-DM pode ser considerado uma metodologia aplicada dentro do ciclo KDD, focando principalmente nas etapas de mineração, avaliação e implantação.

No contexto do processo de KDD, foi aplicada a técnica de mineração de dados por classificação, que tem como objetivo construir um modelo capaz de mapear os descritores de um objeto a um conjunto predefinido de classes, possibilitando a predição da classe de novos exemplos.

Os algoritmos utilizados para essa técnica estão implementados e disponíveis em diversos pacotes de análise de dados. Para esta pesquisa, foi utilizada a ferramenta WEKA, um software livre e gratuito que oferece uma ampla gama de recursos estatísticos para a avaliação e comparação de desempenho de algoritmos de mineração de dados (WEKA WIKI, 2024).

Na ferramenta WEKA foram avaliados os seguintes classificadores: Bayes Net, *Sequential Minimal Optimization* (SMO), J48, Naive Bayes e MultilayerPerceptron, cada um representando abordagens distintas de algoritmos de classificação. O Bayes Net utiliza redes bayesianas, enquanto o Naive Bayes aplica o teorema de Bayes. O J48 constrói árvores de decisão. O SMO é um otimizador para máquinas de vetores de suporte (SVM). Por fim, o MultilayerPerceptron representa uma rede neural artificial com múltiplas camadas.

Esses algoritmos foram aplicados a três cenários distintos: um conjunto de dados não variantes no tempo (*dataset-1*), que se originou do SGA; um conjunto de dados variantes no tempo (*dataset-2*), que se originou no AVA (dados extraídos 30 dias após o início das aulas); e a fusão desses dois conjuntos. As estruturas desses *datasets* são apresentadas, respectivamente, nos Quadros 1 e 2.

Quadro 1 - Estrutura do dataset-1, do SGA, com dados não variantes no tempo

Nome	Descrição
Tipo de ingresso	Vestibular / PROUNI / Transferência interna / Transferência externa / Ingresso como graduados
Tipo de matrícula	Calouro / Reabertura / Renovação / Retorno

Situação no período letivo	Cancelado / Desistente / Desligado / Formado / Matriculado / Trancado / Transferência interna
Período letivo do estudante	<i>numeral</i>
Quantidade de períodos cursados	<i>numeral</i>
Polo de educação a distância	Polos no Brasil: Belém / Belo Horizonte / Brasília / Coronel Fabriciano / Fortaleza / Manaus / Porto Alegre / Recife / Rio de Janeiro / São José dos Campos / São Paulo / Vitória Polos do exterior: Angola / Estados Unidos / Japão
Sexo	M – Masculino / F – Feminino
Estados	Brasil: AM / CE / DF / MG / PA / PE / PI / RJ / RS / SC / SP Exterior: Luanda/Dondo (AO) / Boston (EUA) / Tóquio/Nagoia (JP)
Idade	Abaixo de 20 / 20 a 25 / 26 a 30 / 31 a 35 / 36 a 40 / 41 a 45 / Acima de 45
Estado civil	Casado / Solteiro / Outros / Não informado
Situação final	Evadido / Matriculado

Fonte: Elaboração própria.

Quadro 2 - Estrutura do dataset-2, do AVA, com dados variantes no tempo

Nome	Descrição
Quantidade de acessos à lista de participantes	<i>numeral</i>
Quantidade de acessos ao curso	<i>numeral</i>
Quantidade de acessos ao recurso perfil	<i>numeral</i>
Quantidade de acessos ao quadro de notas	<i>numeral</i>
Quantidade de participações nos fóruns	<i>numeral</i>
Quantidade de acessos aos fóruns	<i>numeral</i>
Quantidade de acessos aos exercícios	<i>numeral</i>
Quantidade de acessos ao plano de ensino	<i>numeral</i>
Quantidade de acessos aos conteúdos	<i>numeral</i>
Classe	Evadido / Matriculado

Fonte: Elaboração própria.

Conforme os Quadros 1 e 2, os dados tratados incluem apenas informações relevantes para a análise educacional, tais como características acadêmicas e comportamentais dos estudantes no AVA e no SGA, garantindo o anonimato dos participantes. Para o treinamento e a avaliação dos algoritmos, os dados são categorizados exclusivamente com o objetivo de identificar padrões relacionados à evasão ou permanência no curso. Dessa forma, nenhuma informação pessoal é exposta, respeitando os princípios da Lei Geral de Proteção de Dados (LGPD).

Após a importação dos dados para uma planilha eletrônica, eles foram submetidos ao WEKA separadamente e, posteriormente, os dados foram unificados. A extração das fontes distintas foi necessária para junção dos dados não variantes no tempo (SGA) e os dados variantes no tempo (AVA).

Os classificadores são aplicados utilizando o método *Cross Validation*, que garante

que cada instância do dado será utilizada pelo menos uma vez para treinamento e testes. Em cada uma dessas vezes, um dos grupos serve para testes e os outros são utilizados para treinamento, garantindo que todos os grupos sejam utilizados uma vez para testes e k-1 vezes para treinamento.

A comparação entre os resultados dos semestres é feita por meio de três métricas de classificação:

- a) Matriz de Confusão, que relaciona classificações corretas com as previsões realizadas para as classes "evadido" e "matriculado", permitindo identificar os verdadeiros positivos e negativos, assim com os falsos positivos e negativos;
- b) precisão, calculada a partir da razão entre previsões verdadeiras positivas e o total de previsões positivas;
- c) acurácia, que representa o percentual de classificações corretas em relação ao total de previsões.

Portanto, diante do processo metodológico estruturado e das análises realizadas, este estudo avança na proposição de soluções baseadas em técnicas de Mineração de Dados Educacionais (EDM) para apoiar a gestão acadêmica. A etapa seguinte deste estudo concentra-se na apresentação dos resultados obtidos com a aplicação dos algoritmos de classificação.

3 RESULTADOS E DISCUSSÕES

A partir dos *datasets* foram realizados seis experimentos, conforme mostrado no Quadro 3.

Quadro 3 - Experimentos da pesquisa

Datasets	Experimento	Status dos Estudantes				
		Matriculado	Formado	Cancelado	Desistente	Trancado
1	1	X		X	X	
	2	X	X	X	X	
2	3	X		X	X	
	4	X	X	X	X	
1 e 2	5	X		X	X	
	6	X	X	X	X	

Fonte: Elaboração própria.

Do SGA foram extraídos inicialmente dez atributos, conforme Quadro 1. Contudo, após submissão dos atributos à função do WEKA que avalia a relevância ou não dos atributos, eles foram reduzidos para três atributos: tipo de matrícula, período letivo do estudante e estado.

Do AVA foram extraídos inicialmente nove atributos, conforme Quadro 2. Entretanto, após a submissão da relevância dos atributos via WEKA, foram utilizados cinco atributos: quantidade de acessos ao curso, quantidade de acesso aos fóruns, quantidade de acesso ao plano de ensino, quantidade de acesso aos conteúdos, quantidade de acesso à lista de participantes.

A partir desses dados, foram realizados seis experimentos com cinco classificadores em cada experimento, totalizando 30 testes, sendo que os melhores resultados de cada experimento são mostrados na Tabela 1.

Tabela 1 - Melhores resultados obtidos nos seis experimentos

Experimento	1	2	3	4	5	6
Classificador	BayesNet/ SMO	Multilayer perception	Naive Bayes	Naive Bayes	Naive Bayes	Naive Bayes
Estudantes formados	Não	Sim	Não	Sim	Não	Sim
Dados	Não variantes no tempo	Não variantes no tempo	Variantes no tempo	Variantes no tempo	Não variantes e variantes no tempo	Não variantes e variantes no tempo
Acurácia	79,7872%	83,6364%	58,5106%	57,2727%	60,6383%	61,8182%
Classificação correta dos evadidos	4 de 21	8 de 21	17 de 21	17 de 21	16 de 21	15 de 21
Precisão	0,1904	0,3809	0,8095	0,8095	0,7619	0,7142

Fonte: Elaboração própria.

Nos experimentos, foram consideradas três métricas de qualidade: matriz de confusão, precisão e acurácia. A análise dos melhores resultados revela a predominância de classificadores bayesianos, especialmente Bayes Net e Naive Bayes. Dos seis classificadores mais eficazes, cinco pertencem a essa categoria, sendo que Naive Bayes aparece quatro vezes e Bayes Net, uma vez.

Outra observação é em relação à inclusão dos estudantes formados no experimento 2, usando os dados “não variantes no tempo”. Tal inclusão apresentou 50% de melhora na quantidade de acertos dos estudantes propensos a evadir, passando de 4

acertos em 21 prováveis no primeiro experimento para 8 acertos em 21 prováveis no segundo experimento. Apesar da melhora, os experimentos mostram a dificuldade na predição da evasão utilizando somente os dados “não variantes no tempo”.

Em relação aos experimentos utilizando dados “variantes no tempo”, os resultados apresentados foram superiores aos experimentos com dados “não variantes no tempo”, sendo que a inclusão dos estudantes formados nos experimentos não apresentou ganho significativo na quantidade de acertos dos estudantes propensos a evadir.

Em relação à acurácia dos resultados, a maior obtida não representou o melhor desempenho para a pesquisa. A maior acurácia foi alcançada no experimento 2, chegando a 83,64%; entretanto, a quantidade de acertos dos estudantes propensos a evadir foi de apenas 8 em um total de 21. Esse resultado demonstra uma limitação da métrica de acurácia nesse contexto, pois ela considera o percentual geral de classificações corretas, sem diferenciar a importância relativa de cada classe. Como a evasão representa uma parcela menor do total de estudantes, um modelo que classifica corretamente a maioria dos alunos como matriculados pode apresentar alta acurácia, mas sem oferecer valor preditivo significativo para a tomada de decisão.

Já nos experimentos 3 e 4, apesar da menor acurácia (58,51% e 57,27%, respectivamente), ambos classificaram corretamente 17 dos 21 estudantes propensos a evadir. Nesse caso, a precisão na identificação do grupo-alvo — os alunos com maior risco de evasão — é mais relevante do que a acurácia geral, pois permite que a instituição intervenha de forma preventiva para reduzir as taxas de abandono. Portanto, um modelo com maior sensibilidade para identificar estudantes em risco, mesmo com menor acurácia geral, oferece maior utilidade para a gestão acadêmica.

Em relação à quantidade de acertos dos estudantes propensos a evadir, dois experimentos se destacaram: os experimentos 3 e 4. Em comum, ambos utilizaram dados “variantes no tempo”, extraídos do AVA, e o classificador Naive Bayes. Os experimentos classificaram corretamente 17 dos 21 estudantes propensos a evadir, com a precisão, métrica de qualidade utilizada para aferir os resultados obtidos, de 0,81%.

Com a união dos dados “não variantes no tempo” e “variantes no tempo”,

esperava-se que os experimentos 5 e 6 apresentassem melhora tanto na classificação dos estudantes propensos a evadir quanto na precisão. Contudo, houve uma discreta piora em relação aos experimentos que utilizaram somente os dados “variantes no tempo”. Com a junção dos dados, foram classificados corretamente 16 dos 21 estudantes propensos a evadir, alcançando uma precisão de 76%.

Com os resultados obtidos, observa-se que foi possível prever os estudantes propensos a evadir, utilizando cinco atributos variantes no tempo, extraídos do AVA, com 30 dias de início das aulas, obtendo uma precisão de 81%.

Os resultados apresentados devem subsidiar os gestores em tomadas de decisão, a partir da elaboração de modelos para tomada de decisão. Esses modelos devem prever ações de acompanhamento dos estudantes durante o semestre letivo, visando: (i) colocar-se à disposição dos estudantes com alguma dificuldade, (ii) solicitar maior empenho dos estudantes que não acessam com frequência o AVA, (iii) lembrar dos prazos das atividades e encontros presenciais avaliativos, (iv) estimular a participação dos estudantes nas atividades acadêmicas, e /ou (vii) envio de informações incentivadoras estimulando a motivação e vínculo do estudante, (viii) provocar os professores para que estimulem os estudantes a participarem das atividades propostas nas disciplinas, (ix) premiar professores com menores índices de evasão.

4 CONSIDERAÇÕES FINAIS

Os resultados deste estudo indicam que a aplicação do processo de KDD nos dados do SGA e do AVA é eficaz na predição de estudantes em risco de evasão. Foram testados algoritmos de classificação mais simples, como Naive Bayes e Bayes Net, que demonstraram boa capacidade preditiva, especialmente ao utilizar dados variantes no tempo. A principal vantagem desses modelos é a sua simplicidade, o que facilita a implementação por parte das universidades, mesmo sem infraestrutura avançada ou equipes especializadas em aprendizado de máquina. Isso permite que instituições de ensino superior adotem estratégias preditivas sem grandes investimentos em tecnologia.

Além disso, a análise dos dados permitiu identificar padrões de comportamento associados à evasão, possibilitando intervenções mais direcionadas por parte da gestão

acadêmica. Isso reforça a importância do uso de técnicas de Mineração de Dados Educacionais (EDM) para o desenvolvimento de estratégias preventivas, contribuindo para a retenção estudantil e para a melhoria da qualidade da EAD.

É importante destacar que este estudo utiliza uma amostra composta por 110 alunos de cursos da área de TI. Os Censos da EaD de diferentes anos apontam que a taxa de evasão no ensino superior varia consideravelmente entre os cursos, o que pode influenciar os resultados de previsões relacionadas à evasão estudantil. Por outro lado, compreendemos que os dados extraídos dos AVA e do SGA possuem natureza genérica, o que pode favorecer sua aplicabilidade a diferentes áreas do conhecimento.

Como proposta para estudos futuros, recomenda-se ampliar a base de dados incluindo um número maior de estudantes, com outros cursos ou áreas e abarcando mais semestres letivos, a fim de aumentar a robustez do modelo e avaliar a possível diferença entre as áreas. Além disso, a utilização de algoritmos mais sofisticados, como Random Forest e redes neurais profundas, pode proporcionar maior precisão na predição da evasão, permitindo a identificação de padrões mais complexos e dinâmicos. Esses modelos avançados podem aprimorar ainda mais a capacidade de análise de dados educacionais, possibilitando intervenções mais eficazes e personalizadas por parte das instituições de ensino.

REFERÊNCIAS

ASSOCIAÇÃO BRASILEIRA DE EDUCAÇÃO A DISTÂNCIA. **Censo EAD.BR 2022**: relatório analítico da aprendizagem a distância no Brasil 2022. Curitiba: InterSaberes, 2024.

Disponível em:

<https://abed.org.br/arquivos/CENSO%20EAD%202022%20PORTUGUES+INGLES.pdf>.

Acesso em: 24 fev. 2025.

FAYYAD, Usama; PIATETSKY-SHAPIO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases. **AI Magazine**, [S. l.], v. 17, n. 3, p. 37, 1996. DOI:

<https://doi.org/10.1609/aimag.v17i3.1230>. Disponível em:

<https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1230>. Acesso em:

18 maio 2025.

GOMES, Josir Cardoso; PIMENTA, Ricardo Medeiros; SCHNEIDER, Marco. Mineração de dados na pesquisa em ciência da informação: desafios e oportunidades. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 20., 2019, Florianópolis. **Anais**

[...]. Florianópolis: UFSC, 2019. Disponível em: <https://conferencias.ufsc.br/index.php/enancib/2019/paper/download/546/517>. Acesso em: 24 fev. 2025.

HOTZ, Nick. **Data science process alliance**: what is CRISP-DM. 2024. Disponível em: <https://www.datascience-pm.com/crisp-dm-2/>. Acesso em: 24 fev. 2025.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA (INEP). **Censo da educação superior: ensino a distância cresce 474% em uma década**. 2022. Disponível em: <https://www.gov.br/inep/pt-br/assuntos/noticias/censo-da-educacao-superior/ensino-a-distancia-cresce-474-em-uma-decada>. Acesso em: 24 fev. 2025.

PIATETSKY, Gregory. **CRISP-DM, still the top methodology for analytics, data mining, or data science projects**. 28 out. 2014. Disponível em: <https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html>. Acesso em: 24 fev. 2025.

WEKA WIKI. **Home**. 2024. Disponível em: <https://waikato.github.io/weka-wiki/>. Acesso em: 24 fev. 2025.



[4.0 Internacional](https://creativecommons.org/licenses/by/4.0/)

Esta licença permite compartilhamento, remixe, adaptação e criação a partir do trabalho, mesmo para fins comerciais, desde que sejam atribuídos créditos ao(s) autor(es). Conteúdos elaborados por terceiros, citados e referenciados nesta obra não são cobertos pela licença.

Como citar este trabalho:

SEPÚLVIDA, Wesley Rodrigues; FERNEDA, Edilson; HELLO, Fernando Antônio; SIQUEIRA, Joyce; KNEIP, Djennifer Hipólito Portugal Amâncio. Uma abordagem preditiva da evasão na educação a distância por meio de técnicas de mineração de dados educacionais. *In*: WORKSHOP DE INFORMAÇÃO DADOS E TECNOLOGIA, 8., 2025, Marília, SP. **Anais [...]**. Marília, SP: Universidade de Marília, 2025. DOI: <https://doi.org/10.22477/viii.widat.220>.



WIDaT 2025

VIII WORKSHOP DE INFORMAÇÃO,
DADOS E TECNOLOGIA

MARÍLIA • SP

Mineração de controvérsias: tensões terminológicas em trabalhos de conclusão de curso

Mining controversies: terminological tensions in undergraduate theses

Benjamin Luiz Franklin¹



<http://lattes.cnpq.br/7192179175216123>

<https://orcid.org/0000-0002-4807-8339>

Resumo

O processo de controle de vocabulário não esconde suas tensões, embates e desafios que caracterizam o desenvolvimento, amadurecimento e normatização das relações entre os atores envolvidos em um domínio. Essas controvérsias podem ser localizadas na evolução de palavras-chave, desde sua utilização inaugural, passando por diferentes grafias, até sua estabilização, enquanto descritores nas publicações científicas. Este trabalho intenta aproximar a mineração de dados à cartografia de controvérsias, associada à Teoria Ator-Rede (TAR), no sentido de colaborar na compreensão do que poderia ser chamado de “mineração de controvérsias”: um processo de localização e mapeamento de pontos de tensão, tradução, negociação e estabilização terminológica em domínios diversos. Após sintetizar os conceitos fundamentais do arcabouço teórico-metodológico da TAR, promovemos um experimento de mineração de dados (e algumas de suas especializações) no acervo de vinte anos dos Trabalhos de Conclusão de Curso (TCCs) do departamento de Ciência da Informação da Universidade Estadual de Londrina (UEL), no sentido de encontrar pontos de controvérsia na utilização das palavras-chave em sua linha do tempo. Após a aplicação de uma função de métrica de controvérsias — ainda em desenvolvimento —, extraímos grupos de termos em conflito e disputa no cenário das pesquisas, espalhados em uma linha do tempo de vinte anos. Os achados sugerem a viabilidade da mineração de controvérsias como um instrumento auxiliar na pesquisa de domínios, destacando a evolução terminológica em documentos respectivos. Futuras pesquisas poderão ampliar o escopo da “Mineração de Controvérsias”, assim como desenvolver a precisão de seus conceitos fundamentais, áreas de aplicação, métricas e funções.

Palavras-chave: mineração de controvérsias; controle de vocabulário; análise de domínio.

Abstract

The process of vocabulary control reveals its tensions, conflicts, and challenges that characterize the development, maturation, and standardization of relationships among actors involved in a domain. These controversies can be traced in the evolution of keywords, from their initial use, through different spellings, to their stabilization as descriptors in scientific publications. This study aims to link data mining to controversy mapping, associated with Actor-Network Theory (ANT), to contribute to understanding what could be called "controversy mining": a process of location and mapping of points of tension, translation, negotiation, and terminological stabilization in various domains. After synthesizing the fundamental concepts of the theoretical-methodological framework of ANT, we conducted a data mining experiment (along with some of its specializations) on a twenty-year collection of undergraduate theses (TCCs) from the Information Science Department at the State University of Londrina (UEL) to find points of controversy in the use of keywords over time. Following the application of a controversy metric function — still under development — we extracted groups of conflicting and disputed terms within the research scenario, spread

¹ Doutor em Engenharia e Gestão de Conhecimento, Universidade Federal de Santa Catarina (UFSC), Florianópolis, Santa Catarina, Brasil. Professor associado do departamento de Ciência da Informação, Universidade Estadual de Londrina (UEL), Londrina, Paraná, Brasil. E-mail: belfra@gmail.com.

across a twenty-year timeline. The findings suggest the feasibility of controversy mining as an auxiliary tool in domain research, highlighting the terminological evolution in respective documents. Future research could expand the scope of "Controversy Mining," as well as develop the precision of its fundamental concepts, areas of application, metrics, and functions.

Keywords: controversy mining; vocabulary control; domain analysis.

1 INTRODUÇÃO

Conforme a síntese de Fujita e Tolare (2019), um vocabulário controlado pode ser entendido como um conjunto organizado de termos e frases usados para indexar conteúdo e recuperar informações de forma consistente. Ele permite que informações semelhantes sejam conectadas e recuperadas usando termos específicos, padronizando a linguagem utilizada em bases de dados, catálogos e repositórios. Sua importância reside na capacidade de melhorar a precisão e eficiência da recuperação de informações, facilitando o acesso a conteúdos relevantes de maneira mais rápida e eficaz, reduzindo a ambiguidade e melhorando a experiência de pesquisa dos usuários.

A construção de um vocabulário controlado não esconde, no entanto, segundo Fujita *et al.* (2023), uma tensão que decorre principalmente da necessidade de balancear a especificidade e abrangência na representação do conhecimento. Por um lado, o vocabulário controlado fornece uma estrutura padronizada que facilita a recuperação de informação, garantindo consistência e precisão; por outro, pode limitar a expressão livre dos pesquisadores, especialistas em seu domínio de estudos, restringindo sua capacidade de descrever nuances e contextos específicos de suas pesquisas. Essa tensão reflete o desafio em atender aos requisitos de sistemas de informação sem comprometer a riqueza e diversidade do conhecimento científico.

Uma vez localizados e tratados, os termos tornam-se "descritores", expressões selecionadas de vocabulários controlados, como tesouros, que são semanticamente relacionados e atuam como instrumentos de controle terminológico (Fujita; Tolare, 2019). Eles são utilizados para realizar a indexação e recuperação das informações de maneira mais precisa. Ajustes entre palavras-chave e descritores de vocabulários controlados incluem, ainda segundo Fujita *et al.* (2023): a equivalência, em que termos correspondentes são utilizados para manter a consistência; especificidade, que ajusta o termo para ser mais específico ou geral; hierarquia, alinhando os termos com a estrutura

hierárquica do vocabulário; e associação, escolhendo termos relacionados para enriquecer a recuperação da informação.

Este trabalho objetiva mapear essas tensões na dinâmica de um domínio, em que as palavras-chave de um documento conflitam com o arcabouço terminológico já utilizado, desafiando o controle do vocabulário e a fruição de seus benefícios. Na próxima seção, descreveremos como podemos entender a dinâmica dessas tensões a partir da Teoria Ator-Rede.

2 PROCEDIMENTOS METODOLÓGICOS

A metodologia empregada neste estudo fundamenta-se na Teoria Ator-Rede (TAR), concebida por Bruno Latour (2005), Michel Callon (1986) e John Law (1992), a partir da década de 1980, como uma abordagem sociotécnica para a compreensão das relações entre atores humanos e não-humanos. A TAR propõe uma perspectiva na qual os atores interagem simetricamente dentro de redes, de modo que suas identidades e capacidades de ação emergem dessas interações. Um aspecto central dessa abordagem é o conceito de “tradução”, definido como o processo por meio do qual os atores negociam, redefinem seus papéis e constroem alianças para estabilizar ou transformar redes (Callon, 1986). No contexto desta pesquisa, a TAR oferece um referencial teórico para examinar as tensões terminológicas identificadas nos Trabalhos de Conclusão de Curso (TCCs), considerando os descritores utilizados como elementos que materializam processos de negociação e estabilização dentro da rede acadêmica (Boullier, 2018).

Para operacionalizar essa análise, utilizou-se a cartografia de controvérsias, metodologia desenvolvida no âmbito da TAR para mapear disputas e conflitos dentro de redes sociotécnicas (Venturini, 2010). O conceito de controvérsia é entendido, conforme Latour (2005), como um momento privilegiado de análise, em que os vínculos entre os atores — humanos e não humanos — tornam-se visíveis. Ao contrário das abordagens que buscam estabilidade ou consenso como objetos principais de estudo, a TAR valoriza as controvérsias justamente porque nelas os elementos que compõem a realidade social entram em disputa, revelando as redes que sustentam fatos, decisões ou instituições.

Assim, a controvérsia não é vista como um ruído ou desvio, mas como uma janela para compreender como o social é continuamente produzido e negociado.

Durante uma controvérsia, diferentes atores mobilizam recursos — argumentos, tecnologias, termos, documentos, dispositivos, autoridades, entre outros — para sustentar posições e construir alianças. A tarefa do pesquisador, nesse contexto, é acompanhar esses movimentos em detalhe, observando como se formam as associações, como se estabilizam certos enunciados e como outros são descartados. Em vez de impor explicações externas, o enfoque está em seguir os atores em sua ação, mapeando as trajetórias pelas quais o social se constitui em meio à controvérsia.

Essa abordagem se caracteriza, dessa forma, por uma investigação empírica detalhada das interações e disputas que emergem em determinados contextos, sem a imposição de categorias analíticas preestabelecidas. A cartografia de controvérsias permite a observação das dinâmicas de poder e da complexidade dos processos de estabilização terminológica, evidenciando como diferentes atores (acadêmicos, instituições e tecnologias) influenciam a constituição do vocabulário acadêmico (Bruno, 2013). Para isso, o presente estudo recorre a métodos computacionais de mineração de dados e algumas de suas especializações, as quais possibilitam a identificação e análise de padrões terminológicos ao longo do tempo, estabelecendo conexões entre o controle do vocabulário e as disputas subjacentes à sua definição (Ryu, 2019).

A mineração de dados é, sintetizando Leetaru (2012), um campo interdisciplinar que engloba técnicas de estatística, aprendizado de máquina e inteligência artificial para extrair informações significativas de grandes conjuntos de dados. A mineração de dados pode ser aplicada de forma ampla, mas, quando orientada a domínios, conforme Liu *et al.* (2023), especifica suas técnicas para se concentrar em áreas particulares ou problemas específicos. Essa abordagem integra conhecimento específico do setor para ajustar os modelos e a interpretação dos dados a um contexto particular. Isso é significativo em campos nos quais o conhecimento especializado é essencial para a compreensão dos dados a partir de um domínio específico ou, sintetizando Smiraglia (2014): uma esfera de conhecimento ou atividade que possui uma base comum de conteúdo e é reconhecida por uma comunidade de indivíduos que compartilham um interesse ou objetivo comum.

A “mineração de ligações” (*link mining*), outra especialização da mineração de dados, é uma técnica que se concentra em identificar e explorar padrões em redes complexas e heterogêneas, formadas por diferentes tipos de nós (como documentos, pessoas ou termos) e arestas, as conexões entre eles (Getoor; Diehl, 2005). A mineração de ligações se destaca por trabalhar com estruturas de grafos, nas quais propriedades como grau, centralidade e conectividade indicam a importância e influência de certos nós e relações. Se essa abordagem específica da mineração de dados se ocupa em evidenciar as relações entre atores em um domínio, compatíveis com as orientações da TAR, a exploração da materialidade textual de um corpus é destacada pela mineração de textos, definida como o processo de extrair conhecimento útil e não trivial a partir de grandes volumes de dados textuais não estruturados (Berry, 2004). A partir de técnicas de aprendizado de máquina, estatística e linguística computacional, busca-se identificar padrões, relações semânticas e estruturas ocultas nos textos.

A análise empírica proposta para este trabalho foi, dessa forma, realizada a partir de um conjunto de especializações do conceito geral de mineração de dados sobre um domínio localizado: um corpus de 291 Trabalhos de Conclusão de Curso defendidos entre 2003 e 2023 no Departamento de Ciência da Informação da Universidade Estadual de Londrina (UEL); e uma materialidade textual única, os textos dos trabalhos. Esses documentos foram selecionados por representar um conjunto consolidado de produções acadêmicas, caracterizadas por uma diversidade de temas, palavras-chave pouco controladas – o que favorece a detecção de controvérsias – e múltiplas abordagens teórico-metodológicas, além de evidenciar materialmente um momento de relações entre diferentes atores – conforme as recomendações da TAR –, como professores, alunos, temas e termos (Universidade Estadual de Londrina, 2017).

Para garantir a consistência dos dados analisados, todas as palavras-chave presentes nos TCCs foram extraídas e organizadas em um banco de dados relacional. Essa etapa foi essencial para estruturar o material e permitir a aplicação das técnicas de mineração supracitadas. A estratégia adotada baseou-se no ciclo de análise de dados descrito por Myatt (2007), que compreende as seguintes etapas: definição do problema, preparação dos dados, implementação da análise e implantação dos resultados.

Na fase de preparação dos dados, foram utilizadas ferramentas específicas para organizar e refinar o corpus textual. Os *softwares Orange, OpenRefine e LibreOffice Base* foram empregados para a limpeza e segmentação dos dados, permitindo a normalização das palavras-chave e a eliminação de possíveis inconsistências. Dessa forma, foram extraídas todas as palavras-chave dos TCCs, sem alterar sua grafia, resultando em um total de 791 palavras-chave. Com essas palavras-chave (PC), foi criada uma matriz de palavras-chave por palavras-chave (PC1 x PC2), totalizando 625.681 possíveis combinações de pares de palavras-chave. Essa matriz possibilitou a identificação de similaridades entre termos, servindo como base para a detecção de variações terminológicas que poderiam refletir controvérsias. Com o objetivo de restringir a análise a termos relevantes, foram considerados apenas aqueles com pelo menos quatro caracteres, a fim de evitar interferências de siglas ou abreviações que pudessem comprometer a interpretação dos resultados.

Para a identificação de variações linguísticas e terminológicas, foram aplicadas duas métricas computacionais: a “distância de Levenshtein” e a função Soundex. A métrica de Levenshtein foi utilizada para medir a similaridade entre palavras-chave, considerando o número mínimo de edições necessárias para transformar um termo em outro (Haldar; Mukhopadhyay, 2011). Foram selecionados termos cuja distância de Levenshtein estivesse entre 2 e 5, resultando em 53 candidatos a palavras-chave controversas. Já a função Soundex, um algoritmo fonético desenvolvido para identificar semelhanças na pronúncia de palavras, permitiu agrupar termos que, embora grafados de forma distinta, apresentavam sons semelhantes (Zobel; Dart, 1996). A interseção dessas duas abordagens resultou na seleção final de 18 palavras-chave que poderiam indicar pontos de tensão no controle do vocabulário.

A última etapa consistiu na análise de redes e na visualização dos dados, utilizando o software Gephi (Bastian; Heymann; Jacomy, 2009) e o algoritmo de modularidade proposto por Blondel *et al.* (2008). Essa técnica permitiu a clusterização dos termos controversos em 7 grupos distintos, revelando padrões de associação e hierarquia dentro do conjunto de palavras-chave analisado. A modularidade foi utilizada para identificar comunidades dentro da rede de termos, destacando agrupamentos de palavras que

compartilhavam características semânticas ou que apresentavam um histórico de disputa terminológica, conforme mostrado na Tabela 1.

Tabela 1 – Palavras-chave controversas e suas funções

IDPC1	Levenshtein	PC1	PC2	Soundex	Tamanho
1	3	Arquivista	Arquivística	A621	10
2	2	Arquivo eclesiástico	Arquivos eclesiásticos	A621	20
3	2	Arquivos eclesiásticos	Arquivo eclesiástico	A621	22
4	3	Arquivística	Arquivista	A621	12
5	3	Biblioteca	Bibliotecária	B143	10
6	2	Biblioteca comunitária	Bibliotecas comunitárias	B143	22
7	2	Bibliotecas comunitárias	Biblioteca comunitária	B143	24
8	3	Bibliotecária	Biblioteca	B143	13
9	3	Bibliotecária	Biblioterapia	B143	13
10	3	Bibliotecário educador	Bibliotecário mediador	B143	22
11	3	Bibliotecário mediador	Bibliotecário educador	B143	22
12	3	Biblioterapia	Bibliotecária	B143	13
13	3	Fases do arquivo	Idades do arquivo	A621	16
14	3	Idades do arquivo	Fases do arquivo	A621	17
15	3	Informação na empresa	Informação para empresa	E516	21
16	3	Informação para empresa	Informação na empresa	E516	23
17	2	Surdez	Surdos	S632	6
18	2	Surdos	Surdez	S632	6

Fonte: Elaboração própria.

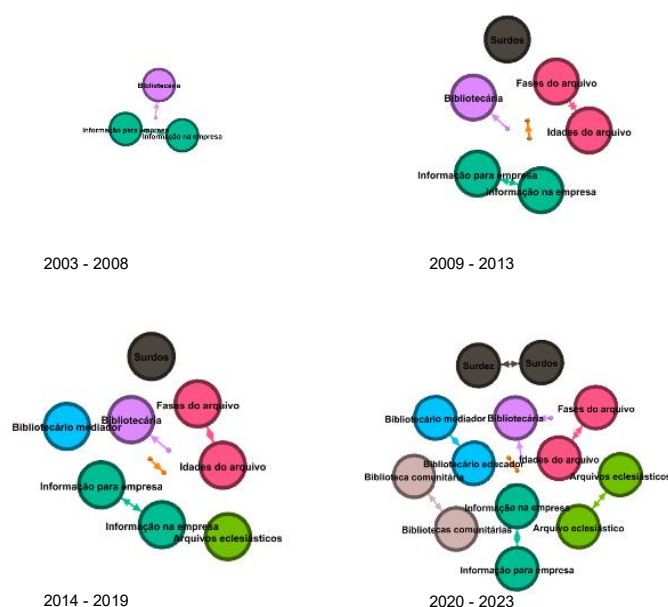
A articulação entre os fundamentos teóricos da Teoria Ator-Rede e as técnicas computacionais de mineração de dados revelou-se produtiva para a identificação e análise das controvérsias terminológicas presentes nos TCCs analisados. A metodologia proposta favoreceu a visualização de padrões e disputas em torno dos termos utilizados,

ao mesmo tempo em que permitiu a construção de uma estrutura analítica voltada à observação dos processos de negociação e estabilização linguística que permeiam a constituição dos vocabulários acadêmicos. Ao final desta etapa, delineou-se um cenário metodológico consistente, capaz de sustentar a análise empírica subsequente, na qual os dados extraídos serão interpretados à luz das redes de controvérsias e das dinâmicas terminológicas que estruturam o domínio em questão.

3 RESULTADOS

Como resultado da aplicação dessas técnicas, na fase final, foram identificados 18 termos semelhantes agrupados em 7 grupos preponderantes. Esses grupos refletem a coesão temática, mesmo a partir de palavras-chave controversas, e podem indicar conceitos correspondentes ou relacionados, agenciados por tensões em torno da estabilização de seu vocabulário. Se visualizadas em um grafo dinâmico, notamos o acúmulo de controvérsias a partir da linha temporal, conforme a Figura 1.

Figura 1 – Dinâmica das controvérsias



Fonte: Elaboração própria.

Os resultados obtidos revelam que as palavras-chave consideradas controversas expressam tensões terminológicas que envolvem variações formais, além de diferenças em como determinados conceitos são compreendidos e empregados ao longo do tempo. A formação de agrupamentos semânticos e a visualização das disputas em rede evidenciam zonas de instabilidade vocabular, sugerindo que o campo pesquisado permanece atravessado por disputas conceituais e acomodações terminológicas. A distribuição temporal dessas controvérsias indica momentos de maior concentração de variações, o que pode sinalizar períodos de transição ou reconfiguração teórica no interior do domínio.

4 CONSIDERAÇÕES FINAIS

Este estudo aborda as tensões terminológicas nos Trabalhos de Conclusão de Curso (TCCs) do departamento de Ciência da Informação da Universidade Estadual de Londrina (UEL), empregando uma metodologia interdisciplinar que integra a Teoria Ator-Rede (TAR) com técnicas de mineração de dados orientadas a domínios, ligações e textos. Ao analisar variações de palavras-chave ao longo de duas décadas, a pesquisa visa capturar a dinâmica de mudança e estabilização terminológica, além de evidenciar as negociações e alinhamentos necessários entre os atores envolvidos — tanto humanos quanto não-humanos — e suas controvérsias terminológicas, vistas como pontos de renegociação em uma rede, conforme sugere a TAR.

A "mineração de controvérsias", como foi delineada neste trabalho, procura evidenciar a evolução da dinâmica do vocabulário acadêmico e os desafios inerentes ao seu controle, destacando pontos de tensão na sua estabilização. Essas localizações colaboram para compreender como os vocabulários controlados podem tanto facilitar quanto obstaculizar a recuperação eficaz de informações em contextos acadêmicos e profissionais, identificando áreas controversas que podem servir como temas para estudos aprofundados, na literatura especializada, ofertando mais um instrumento exploratório para análise de domínios específicos e estudos de vocabulário.

Destacamos, dessa forma, que a mineração de controvérsias apenas evidencia, destaca e aponta pontos de tensão e realinhamento entre traduções entre os atores de

uma rede sociotécnica, sem explicá-los ou justificá-los, tarefas destinadas às pesquisas especializadas. Não obstante, essa abordagem pode auxiliar a construção de problemas teóricos, subsidiando empiricamente caminhos a serem desvendados ou questões de pesquisa que precisem de maiores esclarecimentos.

Pesquisas futuras podem ampliar o escopo de aplicação deste estudo, incluindo a análise de diferentes departamentos ou instituições. Essa abordagem permitiria uma comparação mais abrangente das controvérsias terminológicas em distintos domínios do conhecimento, oferecendo uma visão mais completa das dinâmicas envolvidas na estabilização de vocabulários acadêmicos. Além disso, novos estudos poderiam explorar o desenvolvimento de métodos específicos, incorporando tecnologias de mineração de controvérsias alinhadas aos modelos de linguagem de grande escala (LLMs). Esse aprimoramento possibilitaria a identificação de tensões terminológicas dentro de um escopo semântico e interpretativo mais amplo, tornando o controle de vocabulário mais flexível e adaptável às constantes transformações do conhecimento. Com isso, seria possível aumentar a precisão e a eficácia da indexação e recuperação da informação, além de contribuir significativamente para o avanço da literatura em Ciência da Informação e tecnologias da informação, aprofundando a compreensão sobre a evolução dos domínios científicos e seus desafios terminológicos.

REFERÊNCIAS

BASTIAN, Mathieu; HEYMANN, Sebastien; JACOMY, Mathieu. Gephi: An open source software for exploring and manipulating networks. *In: INTERNATIONAL AAAI CONFERENCE ON WEB AND SOCIAL MEDIA*, 3., 2009, San Jose. **Proceedings [...]**. San Jose: AAAI, 2009. v. 3, n. 1, p. 361-362.

BERRY, Michael W. **Survey of text mining: clustering, classification, and retrieval**. New York: Springer, 2004.

BLONDEL, Vincent D. *et al.* Fast unfolding of communities in large networks. **Journal of Statistical Mechanics: Theory and Experiment**, [S. l.], v. 2008, n. 10, p. P10008, 9 out. 2008.

BOULLIER, Dominique. Médialab stories: how to align actor network theory and digital methods. **Big Data & Society**, [S. l.], v. 5, n. 2, p. 1-13, jul. 2018.

BRANDAU, Ricardo; MONTEIRO, Rosangela; BRAILE, Domingo M. Importância do uso correto dos descritores nos artigos científicos. **Brazilian Journal of Cardiovascular Surgery**, [S. l.], v. 20, p. VII-IX, mar. 2005.

BRUNO, Fernanda. Rastros digitais sob a perspectiva da teoria ator-rede. **Revista FAMECOS**, [S. l.], v. 19, n. 3, p. 681-704, 2013. DOI: 10.15448/1980-3729.2012.3.12893. Disponível em: <https://revistaseletronicas.pucrs.br/ojs/index.php/revistafamecos/article/view/12893>. Acesso em: 13 maio 2024.

CALLON, Michell. Some elements of a sociology of translation: domestication of the scallops and the fishermen of St Brieuc Bay. In: LAW, J. (org.). **Power, action and belief: a new sociology of knowledge?** Londo: [s. n.], 1986, p. 196-223.

CALLON, Michell. Techno-economic networks and irreversibility. **The Sociological Review**, [S. l.], v. 38, n. S1, p. 132-161, 1990.

DIEHL, Christopher P.; GETOOR, Lise. Link mining: a survey. **ACM SIGKDD Explorations Newsletter**, [S. l.], v. 7, n. 2, p. 3-12, 2005.

FUJITA, Mariângela Spotti Lopes *et al.* Atribuição de palavras-chave em trabalhos completos de evento: análise de controle de vocabulário. **Informação & Informação**, Londrina, v. 27, n. 4, p. 1-31, 2023. DOI: 10.5433/1981-8920.2022v27n4p1. Disponível em: <https://ojs.uel.br/revistas/uel/index.php/informacao/article/view/47333>. Acesso em: 4 mar. 2024.

FUJITA, Mariângela Spotti Lopes; TOLARE, Jessica Beatriz. Vocabulários controlados na representação e recuperação da informação em repositórios brasileiros. **Informação & Informação**, Londrina, v. 24, n. 2, p. 93-125, 2019. DOI: 10.5433/1981-8920.2019v24n2p93. Disponível em: <https://ojs.uel.br/revistas/uel/index.php/informacao/article/view/37985>. Acesso em: 4 mar. 2024.

HALDAR, Rishin; MUKHOPADHYAY, Debajyoti. Levenshtein distance technique in dictionary lookup methods: an improved approach. **arXiv**, [S. l.], 6 jan. 2011. Disponível em: <http://arxiv.org/abs/1101.1232>. Acesso em: 22 de maio de 2024.

LATOUR, Bruno. **Reassembling the social: an introduction to actor-network-theory**. Oxford: Oxford University Press, 2005.

LAW, John. After ant: complexity, naming and topology. **The Sociological Review**, v. 47, n. 1, p. 1-14, maio 1999.

LAW, John. Notes on the theory of the actor-network: ordering, strategy, and heterogeneity. **Systems Practice**, [S. l.], v. 5, n. 4, p. 379-393, 1992.

LEETARU, Kalev. **Data mining methods for the content analyst**: an introduction to the computational analysis of content. New York: Routledge, 2012.

LIU, Chuanren *et al.* Recent advances in domain-driven data mining. **International Journal of Data Science and Analytics**, [S. l.], v. 15, n. 1, p. 1-7, 2023.

MYATT, Glenn J.; JOHNSON, Wayne P. **Making sense of data**: a practical approach to exploratory data analysis and data mining. New Jersey: Wiley, 2007.

RYU, S. Using actor network theory for data analysis. *In*: INTERNATIONAL CONFERENCE ON COGNITION AND EXPLORATORY LEARNING IN DIGITAL AGE, 16., 2019, Cagliari. **Anais [...]**. Lisboa: International Association for Development of the Information Society, 2019. Disponível em: <https://eric.ed.gov/?id=ED608653>. Acesso em: 4 jun. 2024.

SMIRAGLIA, Richard P. **The elements of knowledge organization**. [S. l.]: Springer Cham, 2014.

UNIVERSIDADE ESTADUAL DE LONDRINA. **Deliberação – Câmara de Graduação Nº 017/2017**. Aprova o Regulamento do Trabalho de Conclusão do Curso – TCC, do Curso de Biblioteconomia [...]. Londrina: UEL, 2017. Disponível em:

https://sites.uel.br/prograd/wp-content/uploads/documentos/deliberacoes/2017/deliberacao_17_17.pdf

VENTURINI, Tommaso. Diving in magma: how to explore controversies with actor-network theory. **Public understanding of science**, [S. l.], v. 19, n. 3, 258-273, 2010.

ZOBEL, Justin; DART, Philip. Phonetic string matching: lessons from information retrieval. *In*: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 19., 1996, Zurich. **Anais [...]**. New York: Association for Computing Machinery, 1996. Disponível em: <https://dl.acm.org/doi/10.1145/243199.243258>. Acesso em: 22 maio 2024.



[4.0 Internacional](https://creativecommons.org/licenses/by/4.0/)

Esta licença permite compartilhamento, remixe, adaptação e criação a partir do trabalho, mesmo para fins comerciais, desde que sejam atribuídos créditos ao(s) autor(es). Conteúdos elaborados por terceiros, citados e referenciados nesta obra não são cobertos pela licença.

Como citar este trabalho:

FRANKLIN, Benjamin Luiz. *Mineração de Controvérsias: localizando tensões terminológicas em Trabalhos de Conclusão de Curso*. *In*: WORKSHOP DE INFORMAÇÃO DADOS E TECNOLOGIA, 8., 2025, Marília, SP. **Anais [...]**. Marília, SP: Universidade de Marília, 2025. DOI: <https://doi.org/10.22477/viii.widat.222>.



WIDaT 2025

VIII WORKSHOP DE INFORMAÇÃO,
DADOS E TECNOLOGIA

MARÍLIA • SP

Visualização geoespacial da frequência escolar no município de Cachoeiro de Itapemirim/ES

*Geospatial visualization of school attendance in the municipality of Cachoeiro de
Itapemirim/ES*

Henrique Monteiro Cristovão¹

 <http://lattes.cnpq.br/5035919384923489>

 <https://orcid.org/0000-0003-2011-7022>

Douglas Bastos Merêncio²

 <https://lattes.cnpq.br/2944393685367995>

 <https://orcid.org/0009-0001-4234-2135>

Alexandre Alves Valente³

 <https://lattes.cnpq.br/2094903371495518>

 <https://orcid.org/0009-0006-7541-9852>

Dalvan Ribeiro de Almeida⁴

 <http://lattes.cnpq.br/0394943519335156>

 <https://orcid.org/0009-0003-1316-0134>

Camile Pereira de Abreu⁵

 <http://lattes.cnpq.br/6218853669033686>

 <https://orcid.org/0009-0006-0546-8421>

Resumo

Um dos grandes desafios educacionais municipais é lidar com as faltas escolares, pois isso afeta diretamente o desempenho escolar, podendo também levar ao aumento da evasão escolar. O município de Cachoeiro de Itapemirim/ES usa uma estratégia de monitoramento de frequência escolar por meio de coleta diária de dados realizada pelos coordenadores das escolas. Contudo, a secretaria de educação municipal sente falta de uma síntese dessas informações para facilitar o acompanhamento pedagógico e auxiliar em uma tomada de decisão mais ágil. Nesse contexto, o objetivo é oferecer uma visualização de informações sistematizadas, com frequência diária e representação geoespacial sobre a incidência de ausências escolares no município. Com natureza aplicada, abordagem qualitativa, caráter exploratório e descritivo, a pesquisa envolveu

¹ Doutor em Ciência da Informação, Universidade de Brasília (UnB), Brasília, DF, Brasil. Professor, Universidade Federal do Espírito Santo (UFES), Vitória, ES, Brasil. henrique.cristovao@ufes.br.

² Especialista em Cidades Inteligentes, Universidade Federal do Espírito Santo (UFES), Vitória, ES, Brasil. Application developer, NTT Data business solutions, Alegre, ES, Brasil. douglasbm040@gmail.com.

³ Especialista em Cidades Inteligentes, Universidade Federal do Espírito Santo (UFES), Vitória, ES, Brasil. Gerente de Inovação e Cidades Inteligentes, Prefeitura de Cariacica (PMC), Cariacica, ES, Brasil. alexandrevalvesvalente@gmail.com.

⁴ Especialista em Cidades Inteligentes, Universidade Federal do Espírito Santo (UFES), Vitória, ES, Brasil. Analista de Tecnologia da Informação, Universidade Federal do Espírito Santo (UFES), Alegre, ES, Brasil. dalvanra@gmail.com.

⁵ Graduação em Psicologia, Centro Universitário Salesiano (UniSales), Vitória, ES, Brasil. Graduada em Artes Visuais, Universidade Federal do Espírito Santo (UFES), Vitória, ES, Brasil. camileabreudg@gmail.com.

a análise dos dados das frequências escolares, investigações de bases de dados abertas sobre informações geoespaciais do município e a implementação de uma aplicação web em Python. Como resultado principal, um dashboard capaz de oferecer uma visualização de informações sistematizadas, com frequência diária e representação geoespacial sobre a incidência de faltas dos alunos nas turmas das escolas por meio de um mapa de calor que indica um estado de alerta sobre a situação de cada escola. Apesar da aplicação ainda estar em fase de implantação, houve sinalizações importantes sobre o seu potencial em impactos positivos na gestão educacional do município, tal como a redução da evasão escolar, a integração em ações interinstitucionais facilitando a colaboração entre diferentes setores da administração pública, e transparência e participação da comunidade.

Palavras-chave: mapeamento geoespacial; sistemas de informação geográfica; frequência escolar; gestão educacional.

Abstract

One of the major challenges in municipal education is dealing with school absences, as this directly affects school performance and can also lead to an increase in school dropout rates. The municipality of Cachoeiro de Itapemirim/ES uses a strategy to monitor school attendance through daily data collection by school coordinators. However, the municipal education department feels the need for a summary of this information to facilitate pedagogical monitoring and assist in faster decision-making. In this context, the objective is to provide a visualization of systematized information, with daily attendance and geospatial representation information on the incidence of school absences in the municipality. With an applied nature, a qualitative-quantitative approach, and an exploratory and descriptive character, the research involved the analysis of school attendance data, investigations of open databases on geospatial information from the municipality, and the implementation of a web application in Python with publication on a free web server. The main result was a dashboard capable of providing a visualization of systematized information, with daily frequency and geospatial representation information on the incidence of absences in school classes through a heat map that indicates an alert status regarding the situation of each school. Although the application is still in the implementation phase, there were important signs about its potential for positive impacts on the educational management of the municipality, such as the reduction of school dropouts, integration in interinstitutional actions facilitating collaboration between different sectors of public administration, and transparency and community participation.

Keywords: geospatial mapping; geographic information systems; school attendance; educational management.

1 INTRODUÇÃO

A área da educação no Brasil, segundo o Ministério da Educação (MEC), abrange um conjunto de atividades e políticas voltadas ao ensino e a aprendizagem, desde a educação infantil até o ensino superior, envolvendo o planejamento, a execução e a supervisão de ações educacionais em todas as etapas de ensino, visando garantir uma formação integral, ética e cidadã (Brasil, 2024). No contexto municipal, a área da educação é gerida pelas secretarias municipais de educação, que têm um papel essencial na implementação de políticas educacionais e na administração das escolas públicas do ensino básico (educação infantil e ensino fundamental). Cada município possui autonomia para organizar seu sistema de ensino dentro das diretrizes gerais

estabelecidas pelo MEC e pelos estados (Brasil, 2024). Portanto, a administração municipal é fundamental para adaptar as diretrizes nacionais à realidade local, promovendo uma educação que atenda às necessidades específicas da comunidade e contribua para a formação integral dos alunos.

Nesse contexto, um dos desafios significativos e recorrentes nas secretarias de educação municipais, envolvendo a área pedagógica e de ensino, é lidar com a ausência dos alunos nas aulas, pois isso tem afetado diretamente a qualidade da educação e o desempenho escolar. Esse problema, também chamado de absenteísmo escolar, pode levar ao aumento do abandono e posterior evasão escolar, impactando o aprendizado dos estudantes, especialmente na educação básica (Brasil, 2024).

Uma das estratégias de combate a esse problema da ausência escolar, realizada pela Secretaria de Educação do Município de Cachoeiro de Itapemirim do Estado do Espírito Santo, é o monitoramento da frequência escolar pela gestão escolar municipal nas turmas das 84 escolas do município. Apesar da existência de um sistema escolar municipal para registro de faltas escolares pelos professores, muitas vezes ele não é devidamente alimentado e atualizado pelos respectivos docentes responsáveis. Para contornar tal problema, a Secretaria de Educação criou um procedimento paralelo e funcional para registro diário do número de alunos presentes e ausentes que é realizado pela coordenação de cada escola em uma planilha compartilhada em nuvem.

Apesar da secretaria possuir de forma tabulada e organizada os totais das presenças e ausências dos alunos em todas as turmas das escolas municipais, ainda é difícil obter precisamente sínteses estatísticas dos dados de forma ágil. Além disso, não há possibilidade de obter informações resumidas sobre as ocorrências de faltas escolares com representação geoespacial no município ao longo do ano letivo. Devido a essas limitações, fica custoso ou até mesmo inviável agir efetivamente em direção de ações de detecção e prevenção de, por exemplo, possíveis casos de abandono ou evasão escolar.

Nesse contexto, os *dashboards* (ou painéis analíticos) são importantes artefatos agregadores de diversos tipos de visualizações capazes de oferecer interatividade e facilitar a transformação de dados em elementos visuais carregados de semântica. De fato, conforme Chen (2013), as visualizações de informação podem trazer *insights* e

transformar abstrações em elementos concretos e visualmente significativos. Em especial, a utilização de dados geoespaciais na visualização de informação traz possibilidades de construção de conhecimento diferenciadas para um usuário. Sistemas de Informação Geográfica⁶ (SIG) são baseados em dados geoespaciais, amplificam a capacidade do usuário em interpretar dados, quando associados a elementos georreferenciados, e possuem uma vasta aplicabilidade em vários setores da sociedade. GISGeografy (2025) exemplifica 1000 situações de uso de SIG em diversas áreas do conhecimento humano, como agricultura, arqueologia, negócios, mudança climática, segurança pública, educação, saúde política entre outros.

Contudo, Hjørland (2018) destaca que é preciso identificar e conhecer os domínios e grupos sociais aos quais os dados pertencem para que se respeite a sua natureza e seus interesses particulares, conforme prega a epistemologia social.

Alinhado à problemática apresentada, o objetivo geral é oferecer uma visualização de informações sistematizadas, com frequência diária e representação geoespacial sobre a incidência de ausências escolares no município de Cachoeiro de Itapemirim/ES.

Ainda que o município de Cachoeiro de Itapemirim apresente uma elevada taxa de escolarização de crianças de 6 a 14 anos, de 97,2%, conforme o levantamento realizado em 2010⁷, considera-se importante a existência de um serviço de monitoramento da frequência escolar para a prevenção e combate ao problema do absenteísmo escolar com ações e intervenções a serem realizadas de forma precoce juntamente com a área de ensino, coordenação e gestão escolar do município.

A visualização é também considerada um processo de transformação de dados, informação e conhecimento em representações gráficas para apoiar tarefas, tais como análise de dados, exploração de informação, explicação da informação, previsão de tendências, detecção de padrões, descobertas etc. (Zhang, 2008). Dessa forma, com apoio de visualizações georeferenciadas como, por exemplo, um mapa de calor da incidência de

⁶ Sistema de Informação Geográfica (SIG) ou *Geographic Information System* (GIS) é um sistema de hardware, software, informação espacial, procedimentos computacionais e recursos humanos que permite e facilita a análise, gestão ou representação de informação geográfica (Wikipedia, 2024).

⁷ Dados obtidos no site da Prefeitura de Cachoeiro de Itapemirim/ES. Disponível em: <https://www.cachoeiro.es.gov.br/a-cidade/dados-gerais/>.

faltas escolares nas escolas apresentadas em um *dashboard*, tem a capacidade auxiliar na tomada de decisões pela gestão escolar do município.

A presente pesquisa faz parte de um projeto de extensão mais amplo denominado de "Escritórios de Dados Municipais: difusão da cultura de dados para Cidades Inteligentes", da Universidade Federal do Espírito Santo (UFES), funcionando no laboratório de projetos denominado de LabCidades⁸. Os autores do presente trabalho são integrantes do referido projeto de extensão.

2 PROCEDIMENTOS METODOLÓGICOS

A pesquisa classifica-se como natureza aplicada, abordagem quali-quantitativa, e de caráter exploratório e descritivo. O levantamento de requisitos foi realizado por meio de procedimento dialógico, inicialmente de forma presencial na prefeitura de Cachoeiro de Itapemirim e, posteriormente, em reuniões remotas com representantes da secretaria de educação do município.

O trabalho envolveu a análise das planilhas com os dados das frequências escolares, investigações de bases de dados abertas sobre informações geoespaciais do município e a implementação da aplicação web na linguagem *Python* com a sua publicação em um servidor do laboratório.

Todos os dados utilizados na pesquisa, inclusive o código fonte da aplicação desenvolvida, encontra-se disponível no repositório de dados da pesquisa⁹. O acesso à execução da aplicação publicada na web, com dados anonimizados, encontra-se na página de aplicações desenvolvidas¹⁰ do LabCidades.

2.1 Bases de Dados Abertas e Investigação Geoespacial

A utilização de bases de dados abertas foi fundamental para a pesquisa, pois permitiu o acesso a dados essenciais para o mapeamento e análise das escolas municipais. Os dados geoespaciais das escolas foram obtidos no portal do Instituto Jones dos Santos

⁸ LabCidades. Disponível em: <https://labcidades.com.br/>.

⁹ Repositório de dados da pesquisa. Disponível em: <https://github.com/labcidades-ufes/>.

¹⁰ Aplicações desenvolvidas no laboratório. Disponível em: <https://labcidades.ufes.br/>.

Neves¹¹ (IJSN) em conjunto com a Secretaria Estadual de Educação (SEDU). Foi aplicado um filtro sobre os dados baixados para separar apenas as escolas públicas e municipais. Em seguida foi gerado um arquivo GeoJSON¹² para servir de base para o mapa que iria ser apresentado no *dashboard* da aplicação.

Foram também utilizados dados do Sistema Integrado de Bases Geoespaciais do Estado do Espírito Santo (GEOBASES)¹³, que tem foco na padronização dos dados espaciais, facilitando a análise geográfica e a integração de informações oriundas de diferentes fontes (Espírito Santo, 2024b). Essa base foi importante para acessar dados geográficos do município de Cachoeiro de Itapemirim como limites administrativos, rede viária e localização das escolas municipais. Dados coletados pelo *Google Street View*¹⁴ serviram para complementar o mapeamento fornecendo uma visão mais detalhada e contextualizada da localização de algumas escolas para fins de confirmação do seu endereço.

A investigação geoespacial foi realizada por meio de um SIG de código aberto denominado QGIS¹⁵. O uso do QGIS foi essencial para o mapeamento e análise espacial de dados relacionados à área educacional do município, especialmente para o georreferenciamento das escolas do município.

A partir das informações fornecidas pelo *Geobases* e pelo IJSN, o QGIS permitiu o tratamento dos dados vetoriais relacionados às localizações das escolas e aos limites dos bairros. Nesse processo de georreferenciamento, as escolas municipais foram

¹¹ Instituto Jones dos Santos Neves é uma instituição de pesquisa vinculada à Secretaria de Estado de Economia e Planejamento do Estado do Espírito Santo. Seu objetivo principal é fornecer informações que apoiem o planejamento e a execução de políticas públicas, abrangendo aspectos como demografia, economia, meio ambiente e infraestrutura (Espírito Santo, 2024a).

¹² GeoJSON é um formato padrão aberto projetado para representar recursos geográficos simples, juntamente com seus atributos não espaciais. É baseado na *JavaScript Object Notation* (JSON). Os recursos incluem pontos (portanto endereços e locais), sequências de linhas (portanto ruas, rodovias e limites), polígonos (países, províncias, terrenos) e coleções com várias partes desses tipos (Wikipedia, 2019).

¹³ GEOBASES é um sistema que integra diversas bases de dados geoespaciais do ES atuando como uma ferramenta de transparência ativa e uma plataforma que reúne dados geográficos georreferenciados para diversos fins, como planejamento urbano, estudos ambientais e gestão de recursos (Espírito Santo, 2024b). Disponível em: <https://geobases.es.gov.br>.

¹⁴ *Google Street View*. Disponível em: <https://www.google.com/intl/pt-BR/streetview/>.

¹⁵ *Quantum GIS* (QGIS) é uma ferramenta *open source* SIG compatível com *Windows*, *Mac*, *Linux*, mantida por uma comunidade global de desenvolvedores, colaboradores e organizações. Disponível em: <https://qgis.org/>.

apropriadamente localizadas e integrados ao mapa do município por meio do sistema SIRGAS 2000¹⁶ Zona 24 Sul.

2.2 Extração de Dados e Implementação da Aplicação

As tabelas com a frequência escolar são alimentadas diariamente por meio de planilhas compartilhadas com as coordenações locais das escolas por meio de uma nuvem no Google Drive. Elas são organizadas de forma que cada uma das 84 escolas do município é representada por uma planilha contendo diversas tabelas, uma para cada semana de aula, conforme pode ser visto no exemplo da Figura 1. As linhas representam as turmas, as colunas representam os dias da semana, e as células do interior da tabela contém o total de frequência, ausência e percentuais de cada turma em um determinado dia. O nome da escola fica na primeira linha e foi propositalmente omitido na Figura 1.

Figura 1 – Exemplo de tabela de anotações da frequência escolar diária de uma escola, referente a uma semana de aula

SEMANA DATA	2ª FEIRA 10/08			3ª FEIRA 11/08			4ª FEIRA 12/08			5ª FEIRA 13/08			6ª FEIRA 14/08			SEMANA
TURMA	PRESENTES	AUSENTES	PORCENTAGEM DOS PRESENTES	PRESENTES	AUSENTES	PORCENTAGEM DOS PRESENTES	PRESENTES	AUSENTES	PORCENTAGEM DOS PRESENTES	PRESENTES	AUSENTES	PORCENTAGEM DOS PRESENTES	PRESENTES	AUSENTES	PORCENTAGEM DOS PRESENTES	PORCENTAGEM DOS PRESENTES NA SEMANA
MAT I A	10	8	55.56%	10	8	55.56%	10	8	55.56%	12	8	60.00%	10	8	55.56%	57.78%
MAT I B	10	8	55.56%	9	7	56.25%	10	8	55.56%	10	8	55.56%	9	7	56.25%	60.00%
MAT II B	15	2	86.24%	12	5	70.59%	11	5	68.75%	12	4	75.00%	12	4	75.00%	75.00%
MAT II C	14	3	82.35%	11	6	64.71%	13	4	76.47%	13	4	76.47%	14	3	82.35%	76.47%
MAT II D	12	4	75.00%	10	6	60.00%	14	2	87.50%	10	0	100.00%	14	2	87.50%	90.00%
MAT II E	7	10	41.18%	7	10	41.18%	6	8	52.94%	7	10	41.18%	8	12	29.41%	41.18%
MAT III A	9	7	56.25%	13	3	81.25%	10	7	58.82%	12	5	70.59%	14	3	82.35%	69.88%
MAT III B	19	1	95.00%	18	2	90.00%	18	2	90.00%	18	2	90.00%	18	4	80.00%	89.00%
MAT III C	17	2	88.47%	18	3	84.21%	18	3	84.21%	15	4	78.59%	12	7	63.16%	80.00%
MAT III D	16	2	88.89%	17	1	94.44%	17	1	94.44%	18	2	88.89%	12	8	60.00%	89.57%
MAT III E	18	3	84.21%	17	2	88.47%	17	2	88.47%	17	2	88.47%	18	9	66.67%	80.52%
MAT IV A	18	5	78.19%	18	3	85.71%	18	3	85.71%	18	2	90.48%	18	5	78.19%	82.86%
MAT IV B	15	8	71.43%	17	4	80.95%	17	4	80.95%	17	4	80.95%	17	4	80.95%	79.05%
MAT IV C	15	8	71.43%	18	3	85.71%	18	3	85.71%	18	5	78.19%	15	8	71.43%	78.19%
MAT IV D	18	8	72.73%	20	2	90.91%	20	2	90.91%	17	5	77.27%	12	10	54.55%	77.27%
SOMATORIO	207	71	74.46%	219	59	78.78%	218	60	78.42%	217	61	78.06%	197	81	70.66%	76.12%
Taxa de Frequência (média da turma)																

Fonte: Secretaria de Educação de Cachoeiro de Itapemirim/ES (2024).

A implementação da aplicação foi concebida com base no modelo *Extract, Transform, Load* (ETL)¹⁷, utilizando uma abordagem sistemática para transformar dados brutos da frequência escolar advindos de tabelas tipo planilhas em informações organizadas e úteis para o propósito da aplicação.

Para realização da extração de dados na planilha e a organização em outra estrutura, foram selecionadas as variáveis: nome da turma, dia do mês, dia da semana, número de faltosos, número de presentes e percentual de faltosos. Esses elementos

¹⁶ SIRGAS 2000 é o sistema geodésico de referência utilizado no Brasil.

¹⁷ *Extract, Transform, Load* (ETL) refere-se ao processo de extração, transformação e carregamento de dados em um ambiente de data warehouse, sendo essencial para a integração de dados provenientes de diversas fontes, garantindo que eles sejam coletados, limpos, transformados e armazenados de forma eficiente para análise (Vassiliadis, 2009).

foram organizados em uma estrutura de dados formada por tabelas conectadas que facilitaram a análise dos dados promovendo maior clareza e integridade nas informações extraídas.

O algoritmo da aplicação foi projetado para: (i) Reconhecer padrões estruturais nas planilhas, como termos fixos; (ii) Realizar cortes precisos nos dados, extraindo informações relevantes e organizando-as de forma estruturada; e (iii) Automatizar o processo de extração e transformação dos dados, eliminando a necessidade de ajustes manuais e garantindo eficiência e precisão.

A aplicação web foi implementada utilizando o *framework* Dash¹⁸ por sua boa capacidade de integrar bibliotecas *Python* como *Plotly*¹⁹ e *Pandas*²⁰, permitindo o desenvolvimento de *dashboards* dinâmicos e altamente customizáveis. As funcionalidades oferecidas pelo *framework* Dash foram importantes para gerar as interfaces interativas que incluíram gráficos, tabelas e controles deslizantes, permitindo interação pelos usuários.

Para garantir a portabilidade e a reprodutibilidade da solução, foi adotado o uso da plataforma *Docker*²¹ para criar e gerenciar um contêiner para a aplicação desenvolvida. Essa abordagem permitiu a automação do ambiente, facilitando a execução da aplicação tanto em ambientes de desenvolvimento local quanto em produção, garantindo consistência e reduzindo o esforço de configuração manual.

Com a utilização dos elementos tecnológicos citados, a aplicação alcançou um alto nível de flexibilidade e portabilidade, além de simplificação da implantação e boa escalabilidade.

¹⁸ Dash é um *framework open-source* em *Python* amplamente utilizado para criar aplicações web interativas voltadas para visualização de dados. Disponível em: <https://dash.plotly.com/>.

¹⁹ Plotly é uma biblioteca de visualização de dados interativa, comumente utilizada para criar gráficos dinâmicos e informativos. Disponível em: <https://plotly.com/>.

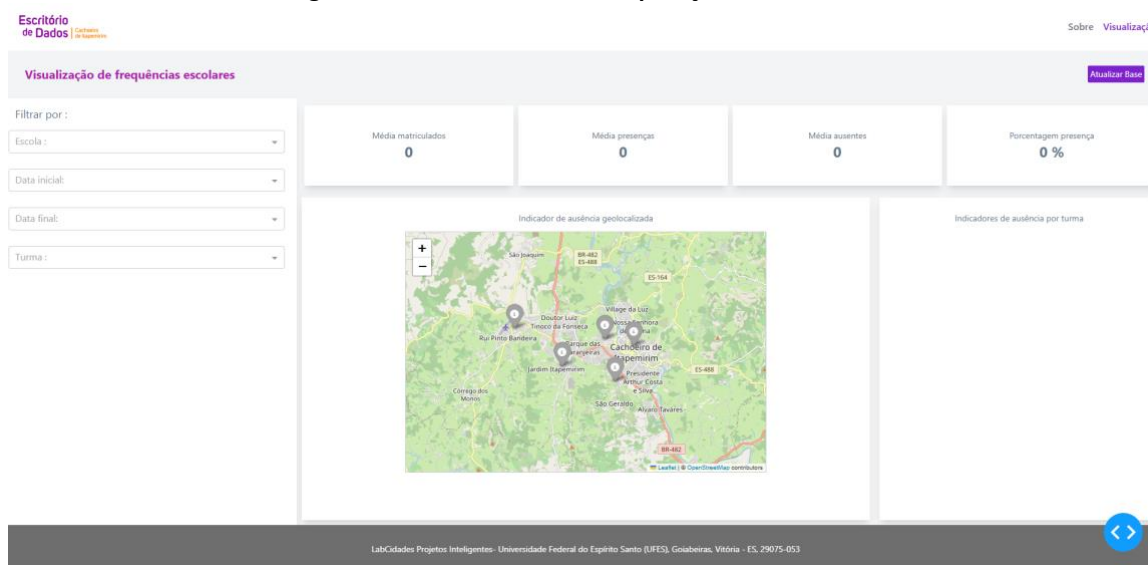
²⁰ Pandas é uma biblioteca de código aberto para *Python* que fornece estruturas de dados poderosas e ferramentas flexíveis para trabalhar com tabelas, séries temporais entre outras. Disponível em: <https://pandas.pydata.org/>.

²¹ Docker é uma plataforma de código aberto que permite criar, implantar e executar aplicativos em contêineres. Ele usa uma tecnologia chamada containerização, que permite empacotar um aplicativo e suas dependências em um contêiner. Disponível em: <https://www.docker.com/>.

3 RESULTADOS E DISCUSSÃO

A visualização de informações sistematizadas como um *dashboard* interativo, com frequência diária e com representação geoespacial sobre a incidência de ausências de alunos nas turmas das escolas por meio de um mapa de calor é apresentada na aplicação desenvolvida, cuja tela de abertura é mostrada na Figura 2. O acesso à execução da aplicação encontra-se na página de aplicações desenvolvidas no LabCidades²².

Figura 2 – Tela de abertura da aplicação desenvolvida



Fonte: Elaboração própria.

Na interface da aplicação, o botão "Atualizar a Base" extrai os dados das 84 planilhas referentes às anotações de frequências das escolas. Em seguida, o usuário pode selecionar as escolas, turmas e período desejado para visualização, no *dashboard*, da síntese dos dados, como mostra o exemplo da Figura 3.

Figura 3 – Síntese de dados de uma determinada escola

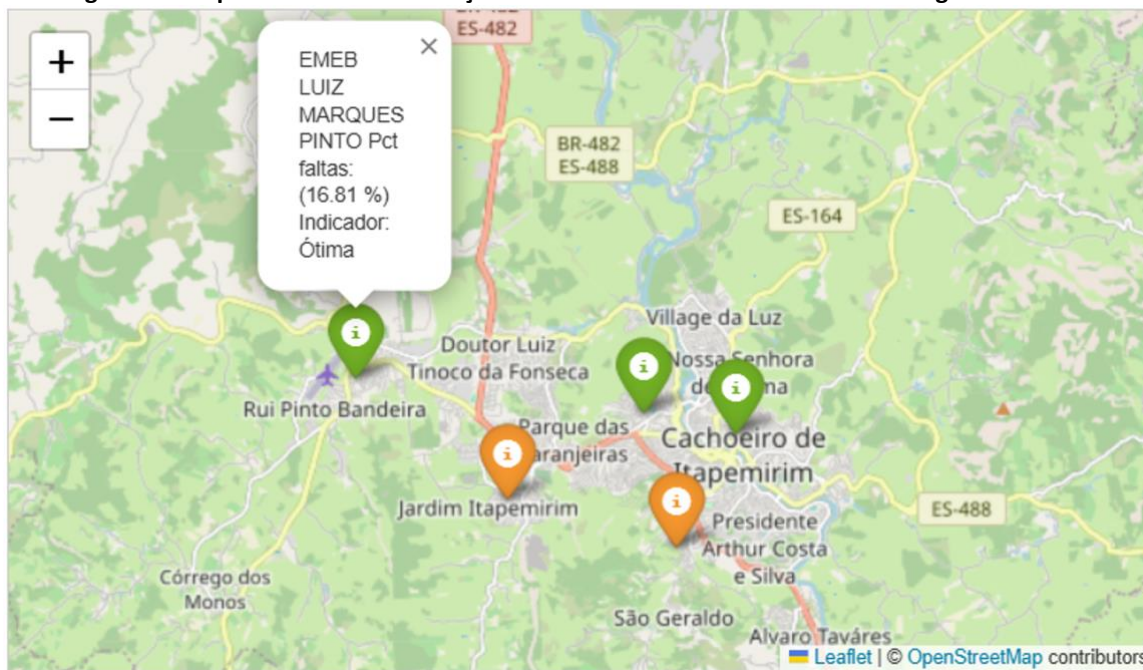


Fonte: Elaboração própria.

²² Aplicações desenvolvidas no laboratório. Disponível em: <https://labcidades.ufes.br/>.

O *dashboard* apresenta também um mapa de calor interativo, no centro da tela, com uma visualização geoespacial das escolas indicando pela cor do marcador o índice de faltas de alunos em cada escola, conforme mostra a Figura 4.

Figura 4 – Mapa de calor com indicação da incidência de faltas de alunos em algumas escolas



Fonte: Elaboração própria.

O mapa utiliza três cores para representar os níveis de absenteísmo: verde, indicando pouco percentual de faltas (até 20%), laranja para percentual mediano de faltas (acima de 20% até 35%), e vermelho para percentual grande (acima de 35%). Os valores foram indicados pela Secretaria de Educação, ainda que de forma experimental, para realização dos testes.

Das 84 escolas do município, apenas cinco aparecem com marcadores, pois como a aplicação ainda está sendo implantada, o acesso aos dados completos das planilhas ainda não foi fornecido pela prefeitura.

A aplicação permite aos gestores da Secretaria de Educação acessar, em tempo real, dados detalhados sobre a presença dos alunos, possibilitando a identificação de padrões de absenteísmo e a localização de áreas críticas do município onde a evasão escolar pode ser mais preocupante. A interface do sistema foi desenvolvida para ser simples e intuitiva, oferecendo filtros que facilitam a visualização segmentada por escola,

turma e diferentes períodos de análise — diário, semanal, mensal ou anual. Contudo, é importante considerar que, segundo Hjørland (2018), os dados nunca são independentes de contexto, mas sempre produzidos a partir de determinadas perspectivas, que contêm critérios próprios de relevância. Assim, a interpretação correta dos *dashboards* é dependente da análise do especialista na Secretaria de Educação do município.

4 CONSIDERAÇÕES FINAIS

Foi desenvolvido um *dashboard* capaz de oferecer uma visualização de informações sistematizadas, com frequência diária e com representação geoespacial sobre a incidência de ausências de alunos nas turmas das escolas por meio de um mapa de calor que indica um estado de alerta sobre a situação de cada escola. Assim, o objetivo geral elencado foi cumprido.

No processo de desenvolvimento da aplicação, destacou-se o QGIS como uma ferramenta de apoio no tratamento e organização de dados geoespaciais, permitindo que as escolas municipais fossem adequadamente georreferenciadas. A linguagem *Python*, munida de bibliotecas específicas, foi importante para viabilizar o desenvolvimento e a publicação da aplicação na web. Destaca-se também que grande parte dos dados foram obtidos a partir de bases de dados abertas como IJSN e GEOBASES.

Apesar da aplicação ainda estar em fase de implantação na prefeitura para uma futura utilização, no momento de sua apresentação à Secretaria de Educação houve sinalizações importantes sobre o seu potencial em impactos positivos na gestão educacional do município, principalmente quanto à redução da evasão escolar, pois permite a identificação precoce de alunos faltosos, possibilitando uma intervenção educacional mais rápida.

Também é importante destacar, que a solução proposta não mudou os processos realizados diariamente pelos profissionais da escola ou da Secretaria de Educação, quanto a anotação de faltas e presenças de alunos. Como consequência disso, a sua implantação se torna mais fácil.

Além da evasão escolar, outros potenciais efeitos positivos foram discutidos, tais como: (i) Aumento da eficiência na gestão escolar, pois o tempo dedicado pelos

professores e coordenadores ao preenchimento de tabelas e relatórios será reduzido, permitindo que os profissionais se concentrem em atividades no âmbito pedagógico propriamente dito; (ii) Tomada de decisão baseada em dados advindos do mapeamento geoespacial das escolas integrados a dados geográficos, pois a Secretaria de Educação terá uma visão mais detalhada sobre as áreas que apresentam maior taxa de evasão; isso permitirá o direcionamento de recursos e políticas públicas de forma mais precisa e estratégica, atendendo às necessidades específicas de cada localidade; (iv) Integração e ações interinstitucionais facilitando a colaboração entre diferentes setores da administração pública como, por exemplo, a área de Assistência Social e Saúde; (v) Transparência e participação da comunidade, uma vez que a aplicação tiver seu uso aberto para a comunidade, envolvendo pais e responsáveis, fortalecendo parceria entre escola e família.

Estes impactos, embora potenciais, indicam possibilidades concretas para melhoria na educação do município, trazendo à tona reflexões sobre a forma como os dados educacionais são coletados, analisados e utilizados em direção ao auxílio da tomada de decisão na área educativa municipal. Além disso, podem promover uma abordagem mais preventiva e colaborativa no enfrentamento dos desafios educacionais, como a evasão escolar.

Como continuidade do estudo indica-se a avaliação e o aperfeiçoamento da ferramenta desenvolvida, analisando-se o impacto real do seu uso e a implementação de novos recursos como, por exemplo, a implantação de alertas automáticos quando ocorrerem aumento repentino de faltas escolares. Por último, uma investigação em direção da transformação da aplicação em um modelo capaz de ser replicado a outras prefeituras.

REFERÊNCIAS

BRASIL. Ministério da Educação. **Portal institucional do MEC**. Brasília: Ministério da Educação, 2024. Disponível em: <https://www.gov.br/mec/pt-br>. Acesso em: 28 fev. 2025.

CHEN, Chaomei. **Mapping scientific frontiers: the quest for knowledge visualization**. London: Springer Science and Business Media, 2013.

ESPÍRITO SANTO (Estado). Secretaria de Estado de Economia e Planejamento. **Instituto Jones dos Santos Neves**. Vitória: Secretaria de Estado de Economia e Planejamento, 2024a. Disponível em: <https://ijsn.es.gov.br/>. Acesso em: 28 fev. 2025.

ESPÍRITO SANTO (Estado). Secretaria de Gestão e Recursos Humanos. **Sistema integrado de Bases Geoespaciais do Estado do Espírito Santo (Geobases)**. Vitória: Secretaria de Gestão e Recursos Humanos, 2024b. Disponível em: <https://geobases.es.gov.br/>. Acesso em: 28 fev. 2025.

GISGEOGRAPHY. GIS Career. **1000 GIS applications and uses**: how GIS is changing the world. 2025. Disponível em: <https://gisgeography.com/gis-applications-uses/>. Acesso em: 28 fev. 2025.

HJØRLAND, Birger. Data (with big data and database semantics). **Knowledge Organization**, [S. l.], v. 45, n. 8, p. 685-708, 2018. Disponível em: <https://www.isko.org/cyclo/data>. Acesso em: 28 fev. 2025.

VASSILIADIS, Panos. A survey of extract–transform–load technology. **International Journal of Data Warehousing and Mining**, [Hershey], v. 5, n. 3, p. 1-27, 2009. Disponível em: <https://ideas.repec.org/a/igg/jdwm00/v5y2009i3p1-27.html>. Acesso em: 28 fev. 2025.

WIKIPEDIA. **GeoJSON**. 2019. Disponível em: <https://pt.wikipedia.org/wiki/GeoJSON>. Acesso em: 28 fev. 2025.

WIKIPEDIA. **Sistema de informação geográfica**. 2024. Disponível em: https://pt.wikipedia.org/wiki/Sistema_de_informa%C3%A7%C3%A3o_geogr%C3%A1fica. Acesso em: 28 fev. 2025.

ZHANG, Jinson. **Visualization for Information Retrieval**. Berlin: Springer, 2008.



[4.0 Internacional](https://creativecommons.org/licenses/by/4.0/)

Esta licença permite compartilhamento, remixe, adaptação e criação a partir do trabalho, mesmo para fins comerciais, desde que sejam atribuídos créditos ao(s) autor(es). Conteúdos elaborados por terceiros, citados e referenciados nesta obra não são cobertos pela licença.

Como citar este trabalho:

CRISTOVÃO, Henrique Monteiro; MERÊNCIO, Douglas Bastos; VALENTE, Alexandre Alves; ALMEIDA, Dalvan Ribeiro; ABREU, Camile Pereira de. Visualização geoespacial da frequência escolar no município de Cachoeiro de Itapemirim/ES. In: WORKSHOP DE INFORMAÇÃO DADOS E TECNOLOGIA, 8., 2025, Marília, SP. **Anais [...]**. Marília, SP: Universidade de Marília, 2025. DOI: <https://doi.org/10.22477/viii.widat.275>.

Neurodados: razões para incidência dos princípios FAIR

Neurodata: reasons for applying FAIR principles

Nathália Pereira Batista Moral¹



<http://lattes.cnpq.br/9948314710971876>



<https://orcid.org/0009-0007-5953-1059>

Silvana Aparecida Borsetti Gregorio Aparecida Vidotti²



<http://lattes.cnpq.br/7390573927636069>



<https://orcid.org/0000-0002-4216-0374>

Resumo

O objetivo da presente pesquisa é o de fomentar debates acerca da necessidade de padrões para representação, interoperabilidade e re(uso) dos neurodados. Valendo-se do método de pesquisa bibliográfico, evidencia-se a necessidade de uma estrutura padronizada para armazenamento e compartilhamento de neurodados, propondo-se, neste cenário, a adoção dos princípios FAIR. As plataformas *OpenNeuro*, *Brain Initiative Data Structure (BIDS)*, *Brain-CODE* e *Neurodata Without Borders* possuem seus próprios padrões de representação. A heterogeneidade de formatos adotados por diferentes laboratórios gera dificuldades para a replicação experimental e limita o uso e reuso eficiente dos dados, comprometendo a progressão da neurociência. A padronização baseada nos princípios FAIR poderia mitigar esses desafios, promovendo maior interoperabilidade, transparência e eficiência na pesquisa. Conclui-se, portanto, ser essencial o aprimoramento de uma linguagem universal para tratamento, armazenamento, localização e interoperabilidade dos neurodados, sendo desejável a aplicação dos princípios FAIR na maior medida possível, para que, assim, a ciência avance em prol da promoção da qualidade de vida e saúde da população.

Palavras-chave: neurodados; dados neurais; princípios FAIR; neurotecnologias.

Abstract

The objective of this research is to foster debates about the need for standards in the representation, interoperability and (re)use of. Using the bibliographic research method, the need for a standardized structure for storing and sharing neurodata is highlighted. In this scenario, we propose the adoption of the FAIR principles. Platforms such as *OpenNeuro*, *Brain Initiative Data Structure (BIDS)*, *Brain-CODE* and *Neurodata Without Borders* have their own representation standards. The heterogeneity of formats adopted by different laboratories creates difficulties for experimental replication and limits the efficient reuse of data, compromising the progression of neuroscience. Standardization based on the FAIR principles could mitigate these challenges, promoting greater interoperability, transparency and efficiency in research. Therefore, we conclude that it is essential to improve a universal language for the treatment,

¹ Doutoranda em Ciência da Informação pela Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP), Marília - São Paulo, Brasil. Oficiala da Justiça Federal (JFSP), Marília - São Paulo, Brasil. E-mail: nathalia.p.batista@unesp.br.

² Doutora em Educação, Universidade Estadual Paulista Júlio de Mesquita Filho (Unesp), Marília, São Paulo, Brasil. Professora do Departamento e do Programa de Pós-graduação em Ciência da Informação, Universidade Estadual Paulista Júlio de Mesquita Filho (Unesp), Marília, São Paulo, Brasil. Coordenadora de Tecnologias Aplicadas (COTEA), Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), Brasília, Distrito Federal, Brasil. E-mail: silvana.vidotti@unesp.br.

storage, location and interoperability of neurodata, and that the application of the FAIR principles is desirable to the greatest extent possible, so that science can advance in favor of promoting the quality of life and health of the population.

Keywords: neurodata; FAIR principles; neurotechnologies.

1 INTRODUÇÃO

A curiosidade sobre o cérebro humano não é recente. Desde o homem pré-histórico, deduções lógicas já nos faziam permitir que traumas cranianos se associavam à perda de memória, convulsões, perda de consciência, dentre outros (Finger, 1994 *apud* Castro; Landeira-Fernandez, 2010). Esta curiosidade progressiva culminou, na década de 1990, em uma verdadeira guerra fria para domínio do mapeamento neurológico e uma melhor compreensão das funções cerebrais.

A Neurociência aplicada na ciência e no mercado se corporifica, essencialmente, em neurotecnologias, as quais produzem neurodados, objeto de estudo do presente trabalho.

Neurodados podem ser definidos como aqueles extraídos, direta ou indiretamente, do sistema nervoso central, neles não incluídas as informações farmacológicas (Brasil, 2022). São, na maior parte dados de pesquisa, frutos da atuação vanguardista de variados laboratórios e do mercado disruptivo, sendo que, cada um destes centros de conhecimento, vale-se de métodos próprios de armazenamento e compartilhamento, que, em grande parte das vezes, não dialogam entre si.

A variedade e a dimensão destes dados desafiam não apenas recursos inovadores e arquivos de armazenamento complexos, mas, principalmente, uma padronização, para que, assim, esforços não sejam despendidos desnecessariamente, compartilhando-se o conhecimento já alcançado por cientistas que nos antecederam.

Por este motivo, tem-se discutido a aplicação dos princípios FAIR para gestão de neurodados, com o propósito de viabilizar a localização, o acesso, a interoperabilidade e (re)uso por instituições científicas, prática estimulada por instituições como a *International Neuroinformatics Coordinating Facility* (INCF), que conta com 18 nações afiliadas (Poline *et al.*, 2022). Trata-se, contudo, de uma grande desafio, em razão complexidade cada vez maior alcançada pelas pesquisas de neurociência, de modo que

buscar esta padronização demanda um esforço internacional conjunto (Poline *et al.*, 2022).

Como resposta, surgem as metodologias *OpenNeuro*, *Brain Imaging Data Structure* (BIDS), *Brain-CODE* e *Neurodata without borders* (NWB), dentre outras, que são tentativas de padronização que permitem compartilhamento e armazenamento de neurodados em um formato único. Diante desse cenário, a adoção dos princípios FAIR é uma solução promissora para propiciar o tratamento, armazenamento, compartilhamento e (re)uso destes dados.

Justifica-se a pesquisa pelo ineditismo nos trabalhos nacionais em Ciência da Informação, tendo-se por objetivo apresentar brevemente o tema, traçando algumas considerações sobre a necessidade da melhor observância dos princípios FAIR sobre os neurodados.

2 PROCEDIMENTOS METODOLÓGICOS

Esta pesquisa adota o método hipotético-dedutivo, estruturado em três etapas principais: (i) levantamento e análise da literatura especializada sobre a gestão de neurodados e os princípios FAIR; (ii) identificação de padrões, convergências e desafios na implementação desses princípios no contexto da neurociência; e (iii) elaboração de uma síntese crítica fundamentada em abordagens comparativas e normativas. A pesquisa é de natureza qualitativa e exploratória, com um delineamento bibliográfico e documental, visando à compreensão aprofundada das implicações técnicas e regulatórias da padronização dos neurodados.

O levantamento bibliográfico foi realizado entre os meses finais de 2024 e início de 2025, com base na revisão de literatura. Foram utilizadas bases de dados científicas indexadas. A busca seguiu critérios rigorosos, incluindo palavras-chave, como *neurodata*, *FAIR principles*, *data interoperability* e *neuroscience*, garantindo a inclusão de estudos pertinentes ao escopo da pesquisa.

Trata-se de investigação atua, em que se compilaram investigações pretéritas na área da Ciência da Informação, as quais foram cruzadas e comparadas com as das áreas

médica e jurídica, a fim de se produzir um conteúdo inédito, explicitado no presente texto.

Além da literatura científica, foram analisados documentos normativos e regulatórios, abrangendo legislações nacionais e internacionais – ex.: Lei Geral de Proteção de Dados (LGPD), *General Data Protection Regulation* (GDPR) e diretrizes da OECD sobre neurotecnologias. A pesquisa documental incluiu jurisprudência relevante e relatórios de organizações científicas e governamentais.

3 NEURODADOS E NEUROTECNOLOGIAS

Apesar de a largada da corrida rumo à melhor compreensão do cérebro humano ter sido dada ainda na década de 1960, foi o Projeto *BRAIN* o responsável por aprofundar o conhecimento científico sobre o sistema nervoso central e seu funcionamento. O *Research Through Advancing Innovative Neurotechnologies (BRAIN)*, deflagrado no ano de 2013 pelo governo americano, é fruto de investimento da Agência de Projetos de Pesquisa Avançada de Defesa do Departamento de Defesa (DARPA) dos EUA neste campo de conhecimento (Farahany, 2023, p. 8, tradução nossa), tendo por finalidade o mapeamento do cérebro humano.

O esforço para aprofundada compreensão das funções cerebrais, contudo, não se limita ao esforço americano, citando-se, ademais, o *China Brain Project*, o projeto japonês *Brain/MINDS*, dentre outros, os quais produzem, constantemente, conhecimento neurocientífico que pode ser intercambiável internacionalmente.

Nacional e internacionalmente, grandes investimentos em iniciativas de neurociência de big data estão sendo realizados (*eHBP*, *US BRAIN*, *ENIGMA*, o *Human Connectome Project* (Elam *et al.*, 2021), o *China Brain Project*, o projeto japonês *Brain/MINDS*, entre outros). Uma miríade de projetos de pesquisa menores, iniciados por pesquisadores, estão continuamente adicionando à nossa base de conhecimento, e pesquisadores individuais estão pensando de forma mais ampla por meio de colaborações estendidas. Com esta avalanche de dados já em curso, existem de facto poucos esforços para coordenar entre comunidades o desenvolvimento de padrões de dados e a representação do conhecimento resultante da investigação experimental (Poline *et al.*, 2022, p. 508, tradução nossa)³.

³ Original: “Is this problem resolving or compounding? Nationally and internationally, large investments in big data neuroscience initiatives are being undertaken (eHBP, US BRAIN, ENIGMA, the Human Connectome Project (Elam *et al.*,

Neurotecnologias são definidas pela Organização para Cooperação e Desenvolvimento Econômico (OCDE) como dispositivos ou procedimentos que acessem, monitorizem, investiguem, avaliem, manipulem e/ou simulem a estrutura e função dos sistemas neuronais de pessoas naturais (Silva, 2022, p. 1141). Alimentam-se de dados como matérias-primas e produzem novos dados, os quais, englobadamente, têm sido denominados de neurodados.

São muitos os usos que se prometem às neurotecnologias, no que se incluem não apenas a cura de doenças, simbiose homem-máquina, entretenimento e exploração pelo mercado, mas também a identificação de estados de humor, o que já é uma realidade em alguns governos orientais.

Segundo a professora e pesquisadora Nita Farahany, na China, maquinistas de trens de alta velocidade no trecho Pequim-Xangai são obrigados a usar dispositivos EEG para monitoramento do estado de vigília. O objetivo é garantir que apenas operadores em estado de alerta estejam em atividade, entre outros usos locais (Farahany, 2023, p. 8, tradução nossa).

Esses exemplos demonstram como tecnologias emergentes, produzem continuamente grandes volumes de dados, os chamados neurodados. Tal volume crescente e especializado de dados gera um interesse significativo por parte da Ciência da Informação, que busca compreender, organizar e gerenciar adequadamente esses dados e informações. Nesse contexto, surgem os princípios FAIR como uma abordagem essencial para assegurar que os dados gerados por essas tecnologias sejam localizáveis, acessíveis e interoperáveis, garantindo uma gestão eficaz e responsável desse tipo de informação.

2021), the China Brain Project, the Japanese Brain/MINDS project, among others). A myriad of smaller, investigatorinitiated research projects are continuously adding to our knowledge base, and individual investigators are thinking more broadly through extended collaborations. With this avalanche of data already underway, there are in fact few efforts to coordinate across communities the development of data standards and knowledge representation resulting from experimental research" (Poline *et al.*, 2022, p. 508).

4 PRINCÍPIOS FAIR

A ciência moderna, de um modo geral, tem sido marcada pela produção e confecção de um volume de dados em larga escala, os quais, compondo informações úteis e preciosas, têm sido amplamente utilizados para tomada de decisões. Assim, faz-se necessário um mecanismo de gestão de dados, em que haja registro de aspectos que viabilizem que a inteligência artificial os localize, acesse e interopere (Vidotti, Torino; Coneglian, 2021, p. 210).

O progresso da ciência é, cada vez mais, conduzido pelo compartilhamento de dados e, para isto facilitar, é crucial a elaboração de um padrão universal (Teeters *et al.*, 2015, p. 629, tradução nossa). Lastreada neste espírito, a comunidade internacional, desde a Conferência Internacional “*Jointly designing the data FAIRPORT*”, de 2014, uniu-se em prol da ampla utilização dos princípios FAIR sobre o armazenamento de dados (Bonetti; Arakaki, 2022, p. 2-3), especialmente para viabilizar um melhor reuso dos científicos.

O avanço da ciência tem sido cada vez mais impulsionado pelo compartilhamento estruturado de dados, sendo essencial a definição de padrões unificados para garantir a localização, acessibilidade, interoperabilidade e (re)uso eficiente destes dados (Teeters *et al.*, 2015, p. 629). Lastreada neste espírito, a comunidade científica internacional consolidou esforços a partir da Conferência Internacional “*Jointly Designing the Data FAIRPORT*”, realizada em 2014, para fomentar a ampla adoção dos princípios FAIR como diretrizes fundamentais para a governança de dados científicos (Henning *et al.*, 2019, p. 177). Esses princípios, voltados para tornar os dados localizáveis, acessíveis, interoperáveis e reutilizáveis, têm sido progressivamente implementados em repositórios de pesquisa, promovendo maior transparência, reprodutibilidade e eficiência na produção do conhecimento científico.

A sigla *FAIR* consiste em um anagrama para as iniciais de “*findable*”, “*accessible*”, “*interoperable*” e “*reusable*”, ou, no português, localizável, acessível, interoperável e reutilizável, e consiste em um mapa para gestores, administradores e controladores de dados, que podem, neles inspirados, averiguar a eficiência do tratamento que dedicam aos dados digitais (Campos; Dias; Souza, 2023, p. 3).

Atentando-se os princípios *FAIR*, otimiza-se não somente a capacidade de organização do homem e da máquina, como também os recursos para se encontrar e se utilizar os dados tratados (Wilkinson *et al.*, 2016). Ainda, segundo Wilkinson *et al.* (2016), os princípios *FAIR* enfatizam o aprimoramento da capacidade das máquinas de localizar e usar os dados automaticamente, além de apoiar seu uso e reuso por indivíduos. O atual ecossistema digital de publicação de dados limita a extração do benefício máximo dos investimentos. O problema não é a falta de tecnologia apropriada, mas sim a falta de um preparo dos dados, para que eles possam ser acionáveis por máquina. Ademais, como nos ensinam Campos, Dias e Souza (2023, p. 2), os princípios *FAIR* direcionam, ainda, a própria elaboração do algoritmo, componente essencial da inteligência artificial que ancora a tomada de decisão pela máquina e pelo homem que se vale das inovações tecnológicas.

É importante frisar que o atendimento dos princípios *FAIR* não se dá de modo binário, não havendo uma única maneira para se atender aos seus princípios, motivo pelo qual se entende haver um espectro *FAIRNESS*, isto é, haver variados graus de a eles se atentar (Higman; Bangert; Jones, 2019 *apud* Bonetti; Arakaki, 2022, p. 3).

Para que se extraia o maior benefício possível, entende-se por ideal a internacionalização dos princípios *FAIR*, já havendo estímulos internacionais para implementação dos dados *FAIR*, como o *FAIR Festival 2021*, o *Open Repositories 2021*, e o *Dataverse Community Meeting*, que demonstram a importância da aplicação dos princípios a nível mundial (Bonetti; Arakaki, p. 3), além da iniciativa *GO FAIR*, ou *Global Open FAIR*, visa disseminar a aplicação dos princípios a todos os países que assim pretendam, para, assim se fazendo, otimizarem-se os serviços de informação.

Essa necessidade de internacionalização dos princípios *FAIR* é especialmente relevante no contexto da Neurociência, área que lida com dados altamente especializados e produzidos em diferentes laboratórios ao redor do mundo. A adoção global de princípios *FAIR* poderia, portanto, promover uma padronização necessária dos neurodados, facilitando sua localização, acesso, interoperabilidade e reuso. Alinhados a esses esforços internacionais, destacam-se iniciativas concretas voltadas à uniformização das práticas em Neurociência, essenciais para superar desafios atuais e impulsionar avanços científicos mais rápidos e consistentes nesse campo.

5 INICIATIVAS DE PADRONIZAÇÃO DE NEURODADOS

Na Neurociência, desde o começo da década de 2010, têm-se reunido esforços para se viabilizar o compartilhamento de dados em larga escala, como aqueles relacionados à neuroimagem (Teeters *et al.*, 2015, p. 629, tradução nossa). Contudo, ainda há o predomínio do tratamento de dados de acordo com a lógica e sistemática individual de cada laboratório, o que dificulta a comparação das técnicas e a replicação de experimentos, um empecilho para o avanço deste ramo da ciência (Teeters *et al.*, 2015, p. 629, tradução nossa). Um plano de gestão de dados, bem como seu armazenamento adequado, é essencial para o reuso, garantindo que sejam novamente utilizados quando necessários para que o progresso científico não parta do zero, mas siga de onde a comunidade acadêmica já tenha alcançado.

A relevância da manutenção dos dados de pesquisa para as gerações futuras é que embasa a preservação digital, compreendida como o 'Conjunto de ações gerenciais e técnicas exigidas para superar as mudanças tecnológicas e a fragilidade dos suportes, garantindo o acesso e a interpretação de documentos digitais pelo tempo que for necessário.' (Conselho Nacional de Arquivos *apud* Torino; Monteiro; Vidotti, 2023, p. 4).

Nos trabalhos científicos, o tradicional é o compartilhamento dos resultados das pesquisas, mas não os dados brutos. Ademais, quando se trata de neurociência, além dos neurodados propriamente ditos, necessário, também a investigação dos metadados, sem os quais, muitas vezes, é impossível a replicação do experimento, motivo pelo qual o tempo e o esforço para a análise dos dados costumam ser elevados (Teeters *et al.*, 2015, p. 629, tradução nossa).

[...] para que os dados sejam Reusable/Reutilizáveis, é importante que tenham metadados exaustivos, precisos e conhecidos pela comunidade, que apresentem uma licença que expresse as condições de uso permitidas pelo titular de direitos autorais dos dados, além dos dados de proveniência (Torino; Monteiro; Vidotti, 2023, p. 7).

Ao citado, acresça-se a importância de um *data paper*, em que constem detalhes de pesquisa, tais como, do equipamento utilizado, formas e conteúdos de coleta, a fim de viabilizar correto reuso destes dados. Portanto, esta conjuntura de complexidade e grande volume de dados torna ainda mais essencial a padronização da forma de armazenamento e busca dos neurodados. “Destaca-se que os princípios FAIR se aplicam

ao tratamento dos dados, dos metadados e das infraestruturas visando maximizar a localização, o acesso, a interoperabilidade e o reuso de dados” (Vidotti; Torino; Coneglian, 2021, p. 210-211).

Seria a neurociência uma disciplina FAIR, onde dados e resultados são encontráveis, acessíveis, interoperáveis e reutilizáveis (Wilkinson *et al.*, 2016)? FAIR é um conjunto de princípios orientadores cada vez mais aceitos para organizar e comunicar os resultados da ciência para que sejam compreensíveis tanto para humanos quanto para máquinas. Em contraste, a comunicação atual de resultados ainda é amplamente baseada em artigos em formato de texto (pdf—html). Os artigos em si são encontráveis e frequentemente acessíveis (graças a iniciativas abertas como o PubMed Central), mas os elementos centrais que levam às conclusões da pesquisa (ou seja, os dados, software, métodos detalhados e resultados completos) raramente são FAIR (Poline *et al.*, 2022, p. 509, tradução nossa)⁴.

É neste cenário que surge uma vasta iniciativa de se tentar padronizar os neurodados, no que se incluem a *OpenNeuro*, a *Brain Imaging Data Structure (BIDS)*, a *Neurodata Without Borders (NWB)* e a *Brain-CODE*. São muitas as iniciativas porque, como, na Sociedade da Informação, estas se traduzem em poder (até mesmo geopolítico), há uma espécie de queda de braço entre laboratórios, porque aqueles que sobre elas têm maior controle ficarão sempre um passo à frente dos demais.

Um exemplo notável de um padrão bem-sucedido é a “Brain Imaging Data Structure”, para dados de ressonância magnética, iniciada em uma reunião do INCF em Stanford, juntamente com o desenvolvimento de muitas ferramentas de análise que dependem do padrão para extrair dados e metadados automaticamente (Gorgolewski *et al.*, 2016, 2017). (...) Um segundo sucesso é o Waxholm Space, um espaço de coordenadas baseado em ressonância magnética 3D para registrar dados em ratos e camundongos em um sistema de coordenadas comum (Johnson *et al.*, 2010; Okamura-Oho *et al.*, 2012; Papp *et al.*, 2014), adotado pelo HBP e E-Brains. O Neurodata Without Borders (NWB) é um padrão emergente para dados fisiológicos que acaba de lançar sua segunda versão e está sendo adotado pela US BRAIN Initiative e outros projetos colaborativos (Rübel *et al.*, 2021). Finalmente, a US BRAIN Initiative está investindo ativamente na criação de novos padrões para dar suporte à neurociência (Poline *et al.*, 2022, p. 510, tradução nossa)⁵.

⁴ Original: “Is neuroscience a FAIR discipline, where data and results are Findable, Accessible, Interoperable and Reusable (Wilkinson *et al.*, 2016)? FAIR is a set of increasingly accepted guiding principles for organizing and communicating the results of science so that they are understandable to both humans and machines. By contrast, current communication of results is still largely based on text (pdf—html) format articles. The articles themselves are findable, and often accessible (thanks to open initiatives like PubMed Central), but the central elements leading to the conclusions of the research (namely the data, software, detailed methods and complete results), are rarely FAIR” (Poline *et al.*, 2022, p. 509).

⁵ Original: “A remarkable example of a successful standard is the “Brain Imaging Data Structure”, for MRI data, started at an INCF meeting at Stanford, along with the development of many analysis tools that rely on the standard to

A OpenNeuro é uma plataforma gratuita, de código aberto, que viabiliza o armazenamento de dados de neuroimagem em formato único, tendo sido lançada no ano de 2011 por Russel Poldrak e Chris Gorgolewski, contando, atualmente, com mais de 55 mil participantes. Trata-se do arquivo do qual se vale a *Brain Initiative*, instituição que tem trabalhado em prol da padronização de neurodados, e que permite o acesso a estes pela própria *web* (OpenNeuro, c2025).

Já a Brain Imaging Data Structure (BIDS) se trata de plataforma de armazenamento inspirada pela OpenNeuro que igualmente se propõe ao armazenamento de dados de neuroimagem, que se vale de um processo de validação e é intercambiável com esta (*Brain Imaging Data Structure*, 2025).

Um sistema fortemente alinhado com os princípios FAIR é o *Brain-CODE* (Behan *et al.*, 2023). Porque poderosos insights podem advir da análise de pesquisas anteriores, a instituição *Ontario Brain Institute* (OBI) envidou esforços para que os neurodados fossem facilmente usados e reusados, culminando em uma plataforma de neuroinformática focada no *FAIRness*, para recolha e partilha generalizada de dados de investigação em neurociência (Behan, 2023).

Uma outra alternativa apresentada é o TemplateFlow, que compreende um repositório baseado em nuvem de modelos de imagens humanas e não humanas — o TemplateFlow Archive, fornecendo o modelo de armazenamento mais adequado àquela pesquisa (Circic *et al.*, 2022, tradução nossa).

Para satisfazer os princípios de interoperabilidade e reutilização, o *TemplateFlow* define um vocabulário para representação inequívoca e legível por máquina de conhecimento e metadados inspirados pela Brain Imaging Data Structure (BIDS). O BIDS prescreve um esquema de nomenclatura de arquivo que compreende uma série de pares de chave-valor (chamados de ‘entidades’). O TemplateFlow adota esse padrão de *design*, pareando cada modelo com uma entidade que o identifica de forma única e persistente. Essa entidade é significada com a chave ‘tpl’ e aceita uma sequência alfanumérica como seu

automatically extract data and metadata (Gorgolewski *et al.*, 2016, 2017). A substantial community has grown around the BIDS standard, with community-led extensions to domains such as magnetoencephalography (Niso *et al.*, 2019) and electroencephalography (Pernet *et al.*, 2019). The community has developed a formal governance procedure for extensions to the protocol, as well as a governance structure with an elected steering group. A second success is the Waxholm Space, a 3D MRI-based coordinate space for registering data in rat and mouse to a common coordinate system (Johnson *et al.*, 2010; Okamura-Oho *et al.*, 2012; Papp *et al.*, 2014), adopted by the HBP and E-Brains. The Neurodata Without Borders (NWB) is an emerging standard for physiological data that has just issued its second version and is seeing uptake in the US BRAIN Initiative and other collaborative projects (Rübel *et al.*, 2021). Finally, the US BRAIN Initiative is actively investing in the creation of new standards to support neuroscience¹⁰ (Poline *et al.*, 2022, p. 510).

valor (por exemplo, 'tpl-MNI152Lin') (Ciric *et al.*, 2022, p. 1569-1570, tradução nossa)⁶.

A plataforma mais divulgada nas pesquisas científicas é a “*Neurodata without borders* (NWB)”, que permite o armazenamento de dados em diferentes formatos, o qual é fruto do trabalho conjunto de neurocientistas com cientistas da computação para a produção de um formato de neurodados padrão, e que acomoda uma enorme heterogeneidade de dados e metadados de neurofisiologia em uma estrutura unificada (Rübel *et al.*, 2022, p. 2). O NWB é capaz de unificar dados e metadados, de diferentes formatos, descrevendo-os e armazenando-os, sendo que dados brutos e processados podem ser armazenados em um mesmo arquivo, o que torna mais fácil o manuseio e a visualização de resultados (Rübel *et al.*, 2022, p. 5).

6 RAZÕES PARA INCIDÊNCIA DOS PRINCÍPIOS FAIR

A expressão “incidência” dos princípios FAIR se justifica em razão: a) do enquadramento dos neurodados como dados; e b) da aplicabilidade dos princípios FAIR àqueles; c) de viabilizar o uso e o reuso de neurodados. Neurodados decorrem de inovações tecnológicas, mas, assim como os demais dados, demandam padronização, aplicação do FAIRness, para que sejam usados, localizados e reusados pelos cientistas e pesquisadores.

A fim de ilustração de um cenário por vir, pontue-se que, no Chile, já há judicialização, datada do ano de 2023, acerca do armazenamento de dados pela empresa Emotiv, cuja retenção de dados, segundo alegado pelo peticionante, poderia viabilizar reidentificação do usuário, vigilância digital e uso de neurodados para fins não autorizados por ele. Trata-se de importante precedente judicial que pode servir de fonte internacional para casos futuros.

⁶ Original: “To satisfy interoperability and reusability principles, TemplateFlow defines a vocabulary for unambiguous and machine-readable representation of knowledge and metadata inspired by the Brain Imaging Data Structure (BIDS)¹³. BIDS prescribes a file-naming scheme comprising a series of key–value pairs (called ‘entities’). TemplateFlow adopts this design pattern, pairing each template with an entity that uniquely and persistently identifies it. This entity is signified with the key ‘tpl-’ and accepts an alphanumeric string as its value (for example, ‘tpl-MNI152Lin’)” (Ciric *et al.*, 2022, p. 1569-1570).

Afinal, a proteção jurídica e constitucional assegurada aos neurodados precisa ser operacionalizada e implementada, o que demanda o auxílio da Ciência da Informação e da Ciência de Dados, sendo a checagem da incidência dos princípios FAIR mais facilmente executável do que a análise casuística de cada sistema diverso, motivo pelo qual a incidência FAIRness deveria ser fomentada não apenas pela iniciativa privada, mas também pelo poder público e pela legislação.

A incidência do FAIRness sobre estes dados pessoais, por viabilizar uma fiscalização mais transparente e ostensiva, iria ao encontro dos princípios jurídicos nacionais e internacionais que prezam pela privacidade e dignidade humana, bens jurídicos tutelados pela Lei Geral de Proteção de Dados (LGPD), havendo, portanto, razões práticas, científicas e jurídicas para incidência do FAIRness sobre os neurodados.

7 CONCLUSÃO

A padronização do formato de tratamento, armazenamento e interoperabilidade dos dados traz vantagens para o uso e reuso dos neurodados por cientistas e pesquisadores, além de tornar mais fluida a fiscalização jurídica sobre estes dados legitimando o uso e o armazenamento pelas instituições de pesquisa e empresas privadas, como já tem sido feito pela Meta, Neuralink ou Emotiv.

Portanto, revela-se essencial o uso de padrões para armazenamento, tratamento e busca por neurodados e desejável a aplicação dos princípios FAIR na maior medida possível, sendo boas alternativas para esta implementação os diferentes *softwares* e plataformas brevemente citadas nesta pesquisa.

REFERÊNCIAS

BEHAN, B.; JEANSON, F.; CHEEMA, H.; ENG, D.; KHIMJI, F.; VACCARINO, A. L.; GEE, T.; EVANS, S. G.; MACPHEE, F. C.; DONG, F.; SHAHNAZARI, S.; SPARKS, A.; MARTENS, E.; LASALANDRA B.; ARNOTT, S. R.; STROTHER, S. C.; JAVADI, M.; DHARSEE, M.; EVANS, K. R.; NYLEN, K.; MIKKELSEN, T. Fair in action: brain-CODE: a neuroscience data sharing platform to accelerate brain research. **Frontiers in Neuroinformatics**, [S. l.], v. 17, may 2023. DOI: <https://10.3389/fninf.2023.1158378> . Disponível em: <https://www.frontiersin.org/journals/neuroinformatics/articles/10.3389/fninf.2023.1158378/full>. Acesso em: 02 mar. 2025.

BONETTI, L. G.; ARAKAKI, A. C. S. Desafios para implementação dos dados FAIR. In: CONGRESSO BRASILEIRO DE BIBLIOTECONOMIA E DOCUMENTAÇÃO, 29., 2022. **Anais [...]**. [online]: FEBAB, 2022. Disponível em: <https://portal.febab.org.br/cbbd2022/article/view/2455>. Acesso em: 09 set. 2024.

BRAIN IMAGING DATA STRUCTURE (BIDS). **Brain imaging data structure 1.10.0**. 2025. Disponível em: <https://bids-specification.readthedocs.io/en/stable/introduction.html>. Acesso em: 02 mar. 2025.

BRASIL. Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais. **Diário Oficial da União**: seção 1, Brasília, DF, n. 157, p. 59, 14 ago. 2018. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 23 fev. 2024.

BRASIL. **Projeto de Lei nº 522/2022**. Modifica a Lei nº 13.709, de 14 de agosto de 2018 (Lei Geral de Proteção de Dados Pessoais), a fim de conceituar dado neural e regulamentar a sua proteção. Brasília, DF: Câmara dos Deputados, 2022. Disponível em: https://www.camara.leg.br/proposicoesWeb/prop_mostrarintegra?codteor=2146384. Acesso em: 9 abr. 2024.

CAMPOS, A. F.; DIAS, G. A.; SOUZA, M. R. F. Princípios FAIR, encontrabilidade e dados: por um diálogo teórico e uma aplicação prática. **Encontros Bibli**, Florianópolis, v. 28, e88060, p. 1-21, 2023. DOI: <https://10.5007/1518-2924.2023.e88060>. Disponível em: <https://periodicos.ufsc.br/index.php/eb/article/view/88060/52528>. Acesso em: 09 set. 2024.

CASTRO, F. S.; LANDEIRA-FERNANDEZ, J. Alma, mente e cérebro na pré-história e nas primeiras civilizações humanas. **Psicologia: Reflexão e Crítica**, Porto Alegre, v. 23, n. 1, p. 141-152, 2010. Disponível em: <https://www.scielo.br/j/prc/a/YD3jJrSVtHgcNjYrJCyScVJ/#>. Acesso em: 26 set. 2024.

CIRIC, R.; THOMPSON, W. H.; LORENZ, R.; GONCALVES, M.; MACNICOL, E. E.; MARKIEWICZ, C. J.; HALCHENKO, Y. O.; GHOSH, S. S.; GORGOLEWSKI, K. J.; POLDRACK, R. A.; ESTEBAN, O. TemplateFlow: FAIR-sharing of multi-scale, multi-species brain models. **Nature Methods**, [S. l.], v. 19, p. 1568-1571, 2022. Disponível em: <https://www.nature.com/articles/s41592-022-01681-2>. Acesso em: 03 mar. 2025.

FARAHANY, N. A. **The battle for your brain**: defending the right to think freely in the age of neurotechnology. New York: St Martin's Press, 2023.

HENNING, P. C.; RIBEIRO, C. J. S.; SALES, L. F.; MOREIRA, J. L. R.; SANTOS, L. O. B. S. Desmistificando os princípios FAIR: conceitos, métricas, tecnologias e aplicações inseridas no ecossistema dos dados FAIR. **Pesquisa Brasileira em Ciência da Informação e Biblioteconomia**, João Pessoa, v. 14, n. 3, p. 175-192, 2019. Disponível em: <https://periodicos.ufpb.br/index.php/pbcib/article/view/46969/27455>. Acesso em: 10 jun. 2024.

OPENNEURO. **Home**. c2025. Disponível em: <https://openneuro.org/>. Acesso em: 02 mar. 2025.

POLINE, J. B.; KENNEDY, D. N.; SOMMER, F. T.; ASCOLI, G. A.; ESSEN, D. C. V.; FERGUSON, A. R.; GRETHE, J. S.; HAWRYLYCZ, M. J.; THOMPSON, P. M.; POLDRACK, R. A.; GHOSH, S. S.; KEATOR, D. B.; ATHEY, T. L.; VOGELSTEIN, J. T.; MAYBERG, H. S.; MARTONE, M. E. Is neuroscience FAIR?: a call for collaborative standardisation for neuroscience data. **Neuroinformatics**, [S. l.], v. 20, p. 507-512, 2022. Disponível em: <https://link.springer.com/article/10.1007/s12021-021-09557-0>. Acesso em: 03 mar. 2025.

RÜBEL, O.; TRITT, A.; LY, R.; DICHTER, B. K.; GHOSH, S.; NIU, L.; BAKER, P.; SOLTESZ, I.; NG, L.; SVOBODA, K.; FRANK, L.; BOUCHARD, K. E. The neurodata without borders ecosystem for neurophysiological data science. **eLife**, [S. l.], v. 11, e78362, 2022. Disponível em: <https://elifesciences.org/articles/78362>. Acesso em: 09 set. 2024.

SILVA, P. M. O direito a permanecer humano: questões éticas das neurotecnologias. **Revista Jurídica Luso Brasileira**, Lisboa, v. 8, n. 1, p. 1441-1459, 2022. Disponível em: https://www.cidp.pt/revistas/rjlb/2022/1/2022_01_1441_1459.pdf. Acesso em: 24 maio 2024.

TEETERS, J. L.; GODFREY, K.; YOUNG, R.; DANG, C.; FRIEDSAM, C.; WARK, B.; ASARI, H.; PERON, S.; LI, N.; PEYRACHE, A.; DENISOV, G.; SIEGLE, J. H.; OLSEN, S. R.; MARTIN, C.; CHUN, M.; TRIPATHY, S.; BLANCHE, T. J.; HARRIS, K.; BUZSAKI, G.; KOCH, C.; MEISTER, M.; SVOBODA, K.; SOMMER, F. T. Neurodata without borders: creating a common data format for neurophysiology. **Neuron**, [S. l.], v. 88, n. 4, p. 629-634, 2015. Disponível em: <https://www.cell.com/action/showPdf?pii=S0896-6273%2815%2900919-8>. Acesso em: 09 set. 2024.

TORINO, E.; MONTEIRO, E. C. S. A.; VIDOTTI, S. A. B. G. Plano de gestão de dados de pesquisa de povos indígenas: considerações acerca dos princípios FAIR e CARE. **Revista Brasileira de Preservação Digital**, Campinas, v. 4, e023007, p. 1-23, 2023. DOI: <https://10.20396/rebpred.v4i00.17936>. Disponível em: <https://econtents.bc.unicamp.br/inpec/index.php/rebpred/article/view/17936>. Acesso em: 23 jan. 2025.

VIDOTTI, S. A. B. G.; TORINO, E.; CONEGLIAN, C. S. #SejaJUSTOeCUIDADOSO: princípios FAIR e CARE na gestão de dados de pesquisa. In: SALES, L. F.; VEIGA, V. S. O.; HENNING, P.; SAYÃO, L. F. (org.). **Princípios FAIR aplicados à gestão de dados de pesquisa**. Rio de Janeiro: IBICT, 2021. p. 209-222. DOI: <https://10.22477/9786589167242.cap15>. Disponível em: https://ridi.ibict.br/bitstream/123456789/1182/2/IBICT_Principios%20FAIR%20aplicado%20a%20gest%C3%A3o%20de%20dados%20de%20pesquisa_2021.pdf. Acesso em: 23 jan. 2025.

WILKINSON, M. D.; DUMONTIER, M.; AALBERSBERG, I. J.; APPLETON, G.; AXTON, M.; BAAK, A.; BLOMBERG, N.; BOITEN, J.-W.; SANTOS, L. B. S.; BOURNE, P. E.; BOUWMAN, J.; A. BROOKES, J.; CLARK, T.; CROSAS, M.; DILLO, I.; DUMON, O.; EDMUNDS, S.; EVELO, C. T.; FINKERS, R.; BELTRAN, A. G.; BELTRAN, A. J. G.; GROTH, P.; GOBLE, C.; GRETHE, J. S.; HERINGA, J.; HOEN, P. A. C 't; HOOFT, R.; KUHN, T.; KOK, R.; KOK, J.; LUSHER, S. J.; MARTONE, M. E.; MONS, A.; PACKER, A. L.; PERSSON, B.; SERRA, P. R.; ROOS, M.; SCHAIK, R. V.; SANSONE, S. A.; SCHULTES, E.; SENGSTAG, T.; SLATER, T.; STRAWN, G.; SWERTZ, M. A.; THOMPSON, M.; LEI, J. V. D.; MULLIGEN, E. V.; VELTEROP, J.; WAAGMEESTER, A.; WITTENBURG, P.; WOLSTENCROFT, K.; ZHAO, J.; MONS, B. The FAIR guiding principles for scientific data management and stewardship. **Scientific Data**, [S. l.], v. 3, e160018, 2016. DOI: <https://10.1038/sdata.2016.18>. Disponível em: <https://www.nature.com/articles/sdata201618>. Acesso em: 09 mar. 2025.



[4.0 Internacional](https://creativecommons.org/licenses/by/4.0/)

Esta licença permite compartilhamento, remixe, adaptação e criação a partir do trabalho, mesmo para fins comerciais, desde que sejam atribuídos créditos ao(s) autor(es). Conteúdos elaborados por terceiros, citados e referenciados nesta obra não são cobertos pela licença.

Como citar este trabalho:

MORAL, Nathália Pereira Batista; VIDOTTI, Silvana Aparecida Borsetti Gregorio. Neurodados: razões para incidência dos princípios FAIR. In: WORKSHOP DE INFORMAÇÃO DADOS E TECNOLOGIA, 8., 2025, Marília, SP. **Anais [...]**. Marília, SP: Universidade de Marília, 2025. DOI: <https://doi.org/10.22477/viii.widat.257>.



WIDaT 2025

VIII WORKSHOP DE INFORMAÇÃO,
DADOS E TECNOLOGIA

MARÍLIA • SP

Desafios da implementação de um Fair Digital Object (FDO): uma proposta de arquitetura para encapsular dados e metadados

Fair Digital Object (FDO) implementation challenges: a framework proposal to encapsulate data and metadata

David Augusto Maduro de Mesquita de Amorim¹



<http://lattes.cnpq.br/6797688041480541>



<https://orcid.org/0009-0005-4830-8295>

Lucas dos Santos Forain^{2,A}



<http://lattes.cnpq.br/5381225685284874>



<https://orcid.org/0009-0005-8627-7629>

Cláudio José Silva Ribeiro^{3,B}



<https://lattes.cnpq.br/1459853686434404>



<https://orcid.org/0000-0002-9571-1707>

Resumo

Apresenta uma investigação sobre o uso de princípios FAIR no contexto do compartilhamento de dados produzidos na área de musicologia. Traz uma contextualização sobre o projeto, trabalhos correlatos desenvolvidos, uso da noção de *Fair Digital Object*, além do Framework associado para sua implementação com enfoque nos identificadores persistentes e representação semântica. Os procedimentos metodológicos foram de cunho qualitativo e exploratório, instrumentalizados pela análise documental da literatura, de modo a reunir os requisitos que foram avaliados. A proposta consiste no encapsulamento dos elementos dentro de um *Container* que poderá ser implementado por meio do uso de uma arquitetura de microserviços, dentro desse *container*, foi organizado o FDO em três camadas principais de metadados e uma camada de dados. Para o resultado foi reunido três modelos de implementação de arquitetura que diferem apenas em relação ao uso de UID. A proposta de arquitetura apresentada visa garantir a conformidade com os princípios FAIR, proporcionando um modelo flexível para diferentes aplicações de FDO. Com esse estudo, espera-se incrementar o debate sobre o uso de FAIR em projetos de pesquisa, em especial no contexto das Humanidades Digitais, com a possibilidade de compartilhar dados para reuso em outras investigações. Conclui-se que a adoção de FAIR e, conseqüentemente, da implementação de FDO, pode ser um aspecto facilitador em projetos tanto no contexto de Instituições de Ciência e Tecnologia, quanto das demais que fazem parte da sociedade organizada.

Palavras-Chave: musicologia; princípios FAIR; Fair Digital Object; arquitetura de software; representação semântica.

¹ Mestrando do Programa de Pós-graduação em Biblioteconomia da Universidade Federal do Estado do Rio de Janeiro (PPGB/Unirio), Rio de Janeiro, RJ, Brasil. davidaugusto1505@edu.unirio.br.

² Graduando em Sistemas de Informação do Centro de Ciências Exatas e Tecnologia da Universidade Federal do Estado do Rio de Janeiro (CCET/Unirio), Rio de Janeiro, RJ, Brasil. lucas.forain@uniriotec.br.

³ Doutor Professor do Programa de Pós-graduação em Biblioteconomia da Universidade Federal do Estado do Rio de Janeiro (PPGB/Unirio), Rio de Janeiro, RJ, Brasil. claudio.ribeiro@unirio.br.

Abstract

This report aims at investigating FAIR principles in the context of sharing the data produced on the musicology fields of study. It provides project contextualization, details related to the work, and utilizes the concept of Fair Digital Object. Also, it brings a framework that emphasizes persistent identifiers and semantic representation in order to implement those FAIR principles. The methodological approach was qualitative and exploratory. They were also put into effect by documentary analysis of the literature, in order to gather the requirements that were evaluated. The proposal consists on encapsulating the elements within a Container that can be implemented through the use of a microservices architecture. Within this container, the FDO was organized into three main metadata layers and a data layer. For the results, three architecture implementation models were gathered, which differ only in relation to the use of UID. The proposed architecture aims to ensure compliance with the FAIR principles, providing a flexible model for different FDO applications. This study expects to increase the debate on the use of FAIR in research projects, especially in the context of Digital Humanities, with the possibility of sharing data to be reused in other investigations. It concluded that the adoption of FAIR and, consequently, the implementation of FDO, can be a facilitating aspect in projects in the context of Science and Technology Institutions and in organized society.

Keywords: musicology; FAIR principles; Fair Digital Object; software architecture; semantic representation.

1 INTRODUÇÃO

Desde o surgimento do movimento do acesso aberto novos desafios têm impulsionado as investigações em Ciência e Tecnologia. As discussões iniciais eram pautadas por acesso às publicações e pelas taxas cobradas pelos editores que eram extremamente prejudiciais aos pesquisadores e respectivas instituições (Apel; Albagli, 2019). Complementarmente, segundo Ribeiro e Almeida (2011) relatam, pode-se afirmar que o movimento de dados abertos e a necessidade de *accountability* no uso de recursos públicos também causaram reflexos diretos para o movimento de transparência na execução das atividades de Ensino e Pesquisa.

Outro ponto motivador foi a necessidade de maior agilidade no atingimento de resultados em investigações. Chen e Jagerhorn (2022) indicam a importância do compartilhamento com forma de melhorar a qualidade nos resultados de investigações.

Nesta direção, os princípios FAIR (*Findable, Accessible, Interoperable, Reusable*) lançaram luz sobre vários aspectos ligados ao compartilhamento. Mas, é possível destacar que as características ligadas aos identificadores persistentes (PID) e metadados, vem provocando debates sobre propostas de arquitetura em objetos FAIR (*Fair Digital Object* - FDO) (FDO, 2025). Vale lembrar que estes temas, PID e metadados, podem ser compreendidos como consequência da aplicação dos princípios F1 e A1, que tratam, respectivamente, de “Metadados com identificador único - *Unique and persistent*

identifier (PID) ou equivalente, para cada conjunto de dados Ex: DOI, ARK, RRID, PID” (Henning *et al.*, 2018, p. 5051) e “Com o identificador do conjunto de dados ou de seus metadados, o usuário deve poder recuperá-los por meio de um protocolo de comunicação padronizado.” (Henning *et al.*, 2018, p. 5051).

Na esteira desta discussão, este trabalho apresenta a proposta de arquitetura para a transformação de *datasets* no contexto do projeto Música em Periódicos Oitocentistas⁴. A proposta é torná-los alinhados aos princípios FAIR, mas seguindo a noção de FDO em conjunto com o framework FDOF (*Fair Digital Object Framework*) (FDOF, 2022).

2 PROCEDIMENTOS METODOLÓGICOS

A presente investigação é um estudo de caso que se caracteriza como de natureza aplicada e exploratória (Braga, 2007). Para a revisão de literatura foram exploradas as fontes: Base de dados Referenciais de Artigos de Periódicos em Ciência da Informação (Brapci), adicionado do Google Acadêmico e acesso ao Portal de Periódicos CAPES (ACM Digital Library e LISTA).

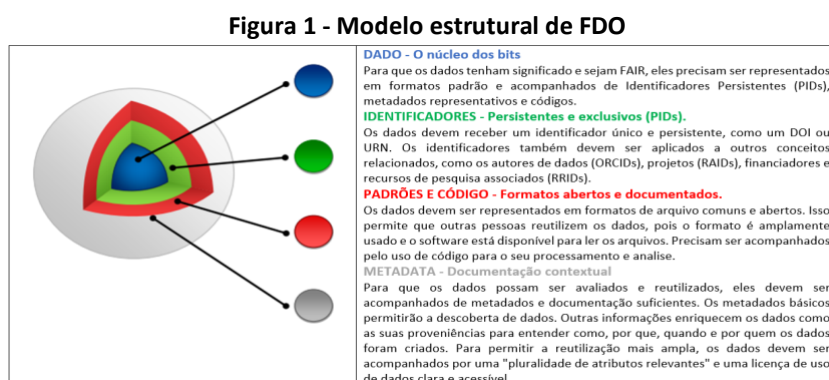
As propostas de uso do FDO foram cotejadas fazendo uso dos pilares: identificadores persistentes e atributos sobre a semântica dos dados. Os autores Lopez Gordillo e Islam (2025), Heinrichs, Hunke e Politze (2025), Santos, Moraes Fonseca e Prince Sales (2025), reunidos na fase de revisão de literatura, trouxeram opções para formular o arcabouço proposto. Por escolha metodológica para a implementação a arquitetura da parte semântica foi detalhada em aspectos descritivos e proveniência.

No contexto do campo empírico, um primeiro recorte necessário esteve ligado ao nível de abstração adotado para o tratamento dos dados. Nesse sentido, a subseção 5.1 explora, por meio dos modelos propostos, os registros que são alvo de publicação.

⁴ Projeto que reúne notícias sobre música do século XIX e está disponível para consulta no endereço <https://musica-sec-xix.unirio.br>. São aproximadamente 4.400 registros coletados desde 2002 de periódicos depositados na Biblioteca Nacional do Brasil.

3 REFERENCIAL TEÓRICO E TRABALHOS CORRELATOS

Esta iniciativa foi gestada a partir de trabalhos correlatos, Ribeiro *et al.* (2023), Ribeiro e Ulhôa (2023; 2024a; 2025a; 2025b, no prelo), para explorar a jornada de implementação de objetos FAIR. Essa implementação parte de uma proposta para arquitetura do FDO de forma a satisfazer o modelo estrutural apresentado na Figura 1. É lícito supor que a arquitetura proposta auxiliará de forma prática a construção de objetos FAIR.

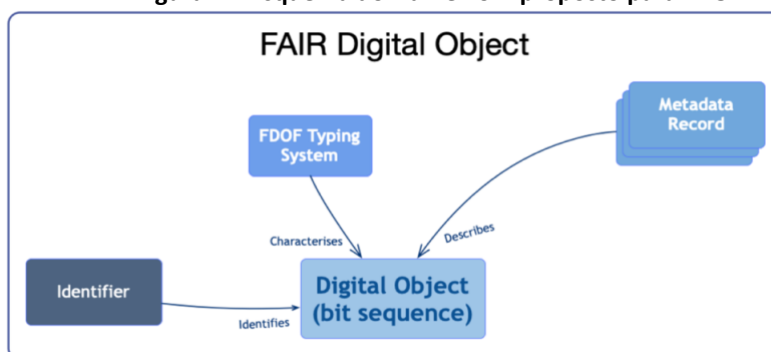


Fonte: Henning *et al.* (2018).

Com o intuito de estudar procedimentos para implementação, o modelo esquemático para encapsulamento apresentado na Figura 1 precisou ser complementado. Nesse sentido, a revisão de literatura sobre arquiteturas para FDO trouxe um conjunto de debates apresentados no *Fair Digital Objects Forum* (FDO, 2025), que, em essência, buscaram explorar: abordagens sobre modelagem semântica; coesão de informações sobre o *dataset*; atendimento aos princípios FAIR; e a busca de alinhamento entre diferentes estruturas de implementação FDO.

De forma complementar, a adoção do FDOF (Figura 2) auxiliou na tarefa de descrever os recursos de dados com a noção de uso de caixa de ferramentas. Os componentes dessa caixa podem agilizar a compreensão e publicação de conjuntos de dados, na medida que “[...] **FAIR Digital Object is defined as a sequence of bits that represents an informational unit and is presented according to the FAIR principles.** [...]” (Santos, 2022, grifo do autor).

Figura 2 – Esquema do framework proposto para FDO



Fonte: Santos (2022).

No *framework* apresentado pela Figura 2 o *identifier* pode ser solucionado por diferentes abordagens, desde atribuição de DOI para conjunto de dados ou então com o uso de serviços em repositórios e *handle system* (Gabriel Junior *et al.*, 2020). De forma análoga, o *metadata record* pode ser atendido por diversos esquemas de metadados e vocabulários (Ribeiro, Santos e Moreira, 2022) que podem ser adotados conforme o campo empírico trabalhado. O Quadro 1 sumariza os metadados definidos em Ribeiro e Ulhôa (2023; 2024b).

Quadro 1: Metadata record

Metadado	Predicado
Data de publicação	dc:date
Data de atualização	dc:date
Fonte	prov:wasGenerateBy
Documentos relacionados	skos:about

Fonte: Recorte de metadados a partir de Ribeiro e Ulhôa (2023; 2024b).

Portanto, explorar o uso de FDOF pode auxiliar a mitigar os riscos advindos de dificuldades na compreensão da relação ternária entre bits-unidade informacional-FAIR, contribuindo para a adoção da Ciência Aberta e inovando especialmente no contexto das Ciências Sociais Aplicadas e das Humanidades Digitais.

4 O CAMPO DE PROVAS

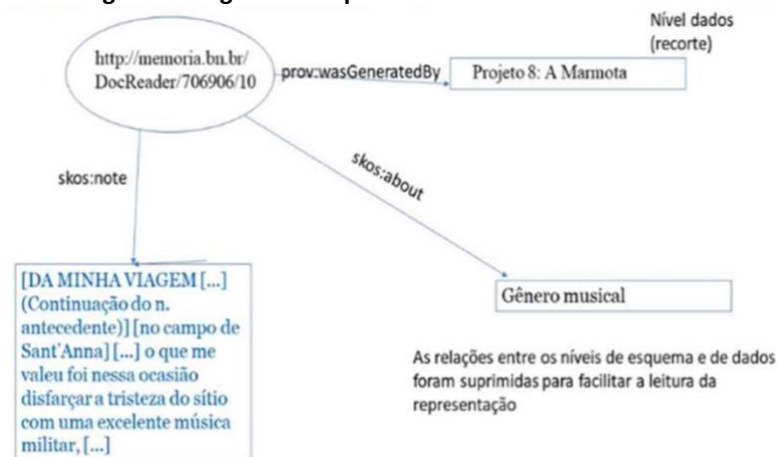
A estrutura de *datasets* para o projeto MPO vem sendo apresentada em diferentes trabalhos (Ribeiro; Ulhoa; 2023; 2024a; 2024b). Os fragmentos apresentados nas Figuras 3 e 4, sintetizam o campo empírico.

Figura 3 – Fragmento esquemático do modelo de classes do projeto MPO



Fonte: Recorte do diagrama com base em Ribeiro e Ulhôa (2023).

Figura 4: Fragmento esquemático do modelo semântico



Fonte: Recorte do diagrama com base em Ribeiro e Ulhôa (2023).

Com um extrato de cerca de 3700 notícias o *dataset* alvo de transformação está publicado no CKAN da Unirio e disponível para download no endereço: <http://dados.unirio.br/dataset/conjunto-de-transcricoes-de-noticias>.

5 A PROPOSTA DE ARQUITETURA PARA GERAÇÃO DO FDO

A implementação do FDO deve seguir princípios de modularidade, escalabilidade e aderência às diretrizes FAIR. A proposta apresentada busca uma abordagem concisa e consistente para a implementação, mas estruturada em diferentes camadas de metadados e dados. Os diagramas para representação seguem o preconizado para o *implementation model* em Jacobsen, Booch e Rumbaugh (1999, p. 273).

A proposta consiste no encapsulamento dos elementos dentro de um *Container* que poderá ser implementado por meio do uso de uma arquitetura de microserviços (Kocher, 2018). Essa proposta pode ser aplicada com o uso da tecnologia *Docker* para garantir isolamento, reprodutibilidade e portabilidade, além de estar disponível em modalidade

Open Source (Rodriguez *et al.*, 2024). Dentro desse container, organizamos o FDO em três camadas principais de metadados e uma camada de dados, conforme proposto abaixo.

- a) **Camada de Metadados de Proveniência:** Esta camada contém informações sobre a origem dos dados, incluindo autor, data de criação, modificações e histórico de versões. Isso garante a rastreabilidade e confiabilidade do FDO (Moreau; Groth, 2013).
- b) **Camada de Metadados de Sintaxe:** Responsável por definir a estrutura dos dados e seus formatos, garantindo a interoperabilidade entre diferentes sistemas (Wilkinson *et al.*, 2016). Essa camada pode incluir esquemas como *JSON Schema*, *XML Schema* ou *RDFS Schema*.
- c) **Camada de Metadados de Identificação Unívoca:** Cada FDO deve ser identificado de maneira única. Esta camada contém um Identificador Único Digital (UID), como um DOI (*Digital Object Identifier*) ou *Handle*, que permite a recuperação dos dados de forma precisa (Schwardmann, 2020; Weigel; Schwardmann; Klump *et al.*, 2020).
- d) **Camada de Dados:** Inclui os próprios dados encapsulados dentro do container, que podem estar em formatos estruturados como JSON, CSV, ou XML. Essa camada é referenciada pela camada de metadados de identificação unívoca (Schwardmann, 2020).

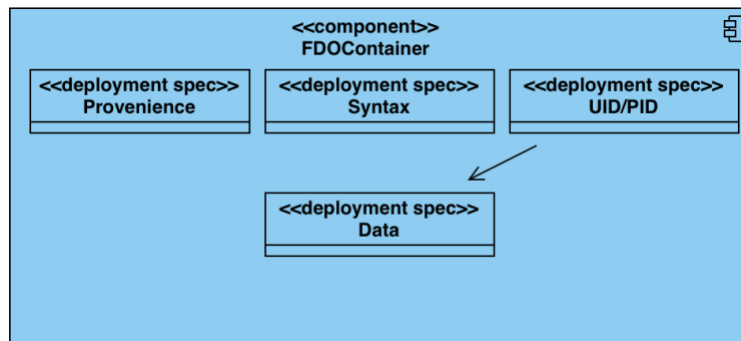
5.1 Modelos de Implementação

Reunimos três modelos de implementação que diferem apenas em relação ao uso de UID. O nível de granularidade para a disponibilização do *dataset* trará reflexos diretos na proposta de implementação de seus identificadores.

5.1.1 Modelo 1: Identificação Única para todo o Conjunto de Dados

A abordagem mais simples consiste em atribuir um único identificador unívoco para todo o conjunto de dados, garantindo um acesso centralizado. Essa estrutura reduz a complexidade e facilita a recuperação de dados agregados (Figura 3).

Figura 3 – Esquema do container FDO com identificador único para o dataset

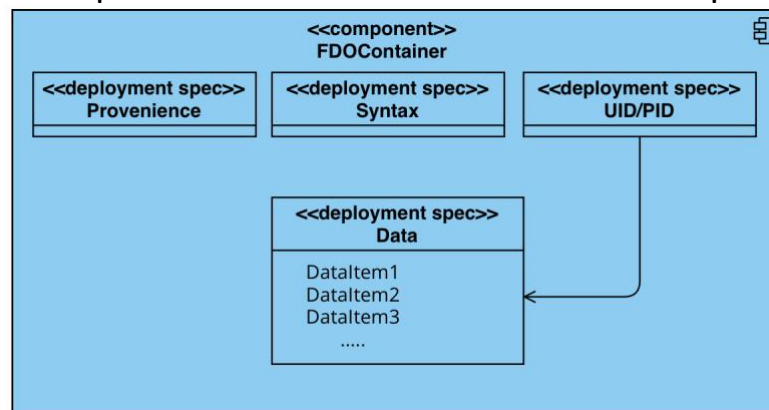


Fonte: Elaboração própria.

5.1.2 Modelo 2: Identificação Unívoca Incremental

Neste modelo, a identificação unívoca é incremental, onde um UDI aponta para a primeira linha dos dados e as seguintes são identificadas pelo incremento desse UDI, semelhante à lógica de alocação de memória. Esse método melhora a granularidade da identificação (Figura 4).

Figura 4 – Esquema do container FDO com identificador incremental para o dataset

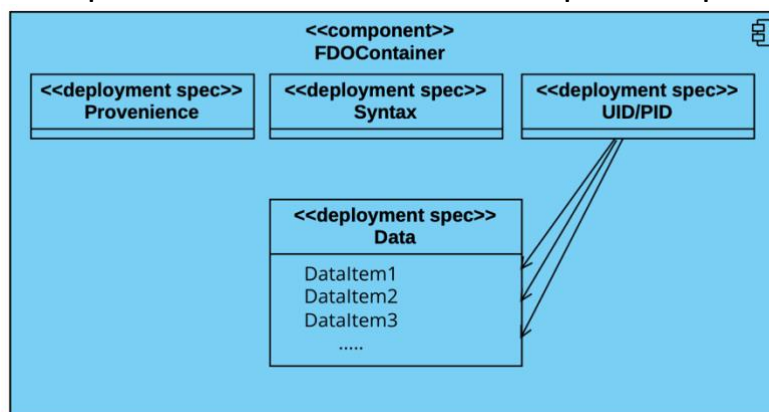


Fonte: Elaboração própria.

5.1.3 Modelo 3: Identificação Unívoca Individual

No terceiro modelo, cada linha do conjunto de dados recebe um UDI único, permitindo que cada item seja referenciado diretamente. Essa abordagem maximiza a reusabilidade e precisão na recuperação de informações (Figura 5).

Figura 5 – Esquema do container FDO com identificador para cada tripla de dados



Fonte: Elaboração própria.

6 DISCUSSÃO

O ponto de partida para a adoção da arquitetura mais adequada está fortemente atrelado ao nível de granularidade do *dataset*. A escolha do modelo de implementação dependerá da necessidade específica a ser atendida para interoperar ou reutilizar os dados.

A arquitetura proposta na seção 5 pressupõe que haverá um PID (*identifier*, mas nomeado na proposta de UID). Esse PID/UID pode ser gerado por um serviço de publicação e que não faz parte do escopo deste trabalho debater. Como observado na seção 2, existem várias maneiras para prover o UID e essas poderão atribuir o identificador único. No protótipo gerado para a publicação dos *datasets* o UID foi gerado pela instância CKAN da Unirio (Ribeiro; Ulhôa, 2024b).

Para as camadas de *Provenience* e *Synthax*: em relação ao esquema de metadados, esses seguem o modelo preconizado pelo CKAN, alinhado ao padrão DCAT (W3C, 201[?]). Foram selecionados para compor o container de armazenamento as propriedades extraídas de vocabulários Web presentes no Quadro 1 e complementadas por meio de consultas realizadas em *Linked Open Vocabularies* (LOV) (LOV, [201?]).

Com os metadados capturados e/ou gerados e todas as conexões entre o objeto-alvo e seus metadados são criadas, a proposta de arquitetura traz o arcabouço de um microserviço SOA⁵ para a criação de FDO. A proposta de arquitetura apresentada visa

⁵ Arquitetura SOA (Service Oriented Architecture) (Ribeiro, 2012).

garantir a conformidade com os princípios FAIR, proporcionando um modelo flexível para diferentes aplicações de FDO.

Vale destacar que o debate sobre caminhos para encapsular as camadas em um FDO e os possíveis desdobramentos para publicá-los em repositórios específicos, mesmo que com o apoio do *Fair Data Point*, pode suscitar a formulação de novos projetos de investigação com caráter interdisciplinar, tanto na área da Ciência da Informação quanto no campo da Ciência da Computação.

7 CONSIDERAÇÕES FINAIS

Como visto na introdução deste relato, o compartilhamento de dados é elemento fundamental para a evolução das pesquisas em diversos campos do conhecimento. Nesse sentido, a adoção de FAIR e, conseqüentemente, da implementação de FDO, pode ser um aspecto facilitador em projetos tanto no contexto de Instituições de Ciência e Tecnologia, quanto da sociedade que pode estar organizada em torno da produção de conhecimento.

Por fim, além trazer a experiência prática e inovadora no compartilhamento de dados de pesquisa, esta pesquisa se alinha com as recomendações de implementação com a iniciativa GoFair quando recomenda o uso de três pilares essenciais para o processo de implementação de infraestrutura em FAIR – Data, Tools and Services (IFDS, [s.d]). Com a apresentação deste estudo de caso espera-se incrementar o debate sobre o uso de FAIR em projetos de pesquisa, em especial no contexto das Humanidades Digitais.

REFERÊNCIAS

APPEL, André Luiz; ALBAGLI, Sarita. Acesso aberto em questão: novas agendas e desafios. **Informação & Sociedade**, João Pessoa, v. 29, n. 4, p. 187, 2019.

BRAGA, Katia Soares. Aspectos relevantes para a seleção de metodologia adequada à pesquisa social em Ciência da Informação. *In*: MUELLER, Suzana Pinheiro Machado (ed.). **Métodos para pesquisa em Ciência da Informação**. Brasília: Ed. Thesaurus, 2007. p. 17-38

CHEN, Xiaoli; JAGERHORN, Martin. Implementing FAIR workflows along the research lifecycle. **Procedia Computer Science**, v. 211, p. 83-92, 2022.

FDO. International FAIR Digital Objects Implementation Summit 2024. **Open Conference Proceedings (OCP)**, Berlin, v. 5, 2025. DOI: <https://doi.org/10.52825/ocp.v5i>. Disponível em: <https://www.tib-op.org/ojs/index.php/ocp/issue/view/24>. Acesso em: 15 abr. 2025.

FDOF. **Fair Digital Object Framework**. 2022. Disponível em: <https://fairdigitalobjectframework.org>. Acesso em: 14 abr. 2025.

GABRIEL JUNIOR, Rene Faustino *et al.* **Acesso aberto a dados de pesquisa no Brasil: identificadores persistentes para dados de pesquisa**. Porto Alegre: CEDAP UFRGS, 2020. Disponível em: <http://hdl.handle.net/20.500.11959/1258>. Acesso em: 15 abr. 2025.

HEINRICHS, Benedikt; HUNKE, Sirieam; POLITZE, Marius. Combining FAIR Digital Object Implementation Concepts in a Real-Life Application. **Open Conference Proceedings (OCP)**, Berlin, v. 5, 2025. DOI: <https://doi.org/10.52825/ocp.v5i.1173>. Disponível em: <https://www.tib-op.org/ojs/index.php/ocp/article/view/1173>. Acesso em: 15 abr. 2025.

HENNING, Patrícia Corrêa *et al.* Desmistificando os princípios FAIR: conceitos, métricas, tecnologias e aplicações inseridas no Ecossistema dos dados FAIR. *In*: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMACÃO, 19., 2018, Marília. **Anais [...]**. Marília, SP: Unesp, 2018.

IFDS. **Internet of Fair & Data and Services**. Disponível em: <https://www.go-fair.org/resources/internet-fair-data-services/>. Acesso em: 15 abr. 2025.

JACOBSON, Ivar; BOOCH, Grady; RUMBAUGH, James. **The unified software development process**. [S. l.]: Addison-Wesley, 1999. v. 1.

KOCHER, Parminder Singh. **Microservices and containers**. [S. l.]: Addison-Wesley Professional, 2018.

LOPEZ GORDILLO, Julian; ISLAM, Sharif. FDO as an Interoperability Framework for the Biodiversity Digital Twin Project. **Open Conference Proceedings (OCP)**, Berlin, v. 5, 2025. DOI: <https://doi.org/10.52825/ocp.v5i.1190>. Disponível em: <https://www.tib-op.org/ojs/index.php/ocp/article/view/1190>. Acesso em: 15 abr. 2025.

LOV. **Linked Data Vocabularies**. [201?]. Disponível em: <https://lov.linkeddata.es/dataset/lov>. Acesso em: 15 abr. 2025.

MOREAU, Luc; GROTH, Paul. **Provenance: an introduction to PROV**. [S. l.]: Morgan & Claypool Publishers, 2013.

RIBEIRO, Claudio José Silva *et al.* Knowledge Organization no processo de FAIRificação de Datasets: estruturando a semântica e interligando as notícias do banco de dados de periódicos musicais oitocentistas. *In*: TOGNOLI, Natália Bolfarini; ALBUQUERQUE, Ana Cristina de; CERVANTES, Brígida Maria Nogueira (org.). **Organização e do conhecimento em diferentes contextos: desafios e perspectivas na era da datificação**. Londrina: UEL, 2023. p. 363-372. Disponível em: https://isko.org.br/wp-content/uploads/2023/06/livro-isko-Brasil_23.pdf. Acesso em: 15 abr. 2025.

RIBEIRO, Claudio Jose Silva; ALMEIDA, Reinado Figueiredo de. Dados abertos governamentais (open government data): instrumento para exercício de cidadania pela sociedade. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 7., 2011, Brasília. **Anais [...]**. Brasília: UnB, 2011. Disponível em: <http://enancib.ibict.br/index.php/enancib/enancibXII/paper/view/881>. Acesso em: 15 abr. 2025.

RIBEIRO, Cláudio José Silva; SANTOS, Luiz Olavo Bonino da Silva; MOREIRA, João Luiz Rebelo. Princípios FAIR e interoperabilidade: reprodutibilidade com o uso de modelagem conceitual orientada a ontologias. **Informação & Informação**, Londrina, v. 25, n. 3, p. 28-51, out. 2020. DOI: <https://doi.org/10.5433/1981-8920.2020v25n3p28>. Disponível em: <https://ojs.uel.br/revistas/uel/index.php/informacao/article/view/41874>. Acesso em: 15 abr. 2025.

RIBEIRO, Claudio José Silva; ULHÔA, Martha Tupinambá de. A publicação de dados em humanidades digitais com Fair: datasets de pesquisa do projeto MPO compartilhados no CKAN da Unirio. **Revista Ci.Inf.**, Brasília, v. 54, n. 3, 2025a. DOI: <https://doi.org/10.18225/ci.inf.v53i3.7222>. Disponível em: <https://revista.ibict.br/ciinf/article/view/7222/6895>. Acesso em: 15 abr. 2025.

RIBEIRO, Claudio José Silva; ULHÔA, Martha Tupinambá de. El uso compartido de conjuntos de datos de investigación del proyecto Música en Periódicos Ochocentistas: un prototipo con el uso de la solución CKAN. **Revista EDICIC**, [S. l.], v. 3, n. 3, p. 1-18, 2023. DOI: <https://doi.org/10.62758/re.v3i3.214>. Disponível em: <https://ojs.edicic.org/index.php/revistaedicic/article/view/214>. Acesso em: 15 abr. 2025.

RIBEIRO, Claudio José Silva; ULHÔA, Martha Tupinambá de. Linked data com Wikipédia e Wikidata: reduzindo os silos de informação na Web com notícias sobre Música em Periódicos Oitocentistas (MPO). In: PESCHANSKI, João Alexandre; JURNO, Amanda Chevtchouk (org.). **A Wikimedia no Brasil: o poder e os desafios do conhecimento livre**. Salvador: EDUFBA, 2025b. No prelo.

RIBEIRO, Claudio José Silva; ULHÔA, Martha Tupinambá de. Reúso de dados de pesquisa em humanidades digitais: investigando textos em notícias do projeto Música em Periódicos Oitocentistas. In: SILVA, Carlos Guardado da; REVEZ, Jorge; CORUJO, Luís (ed.). (2024). **Diálogos na Ciência da Informação = Diálogos en Ciencia de la Información**: atas do XIV EDICIC. Lisboa: Universidade de Lisboa, 2024a. Disponível em: <http://hdl.handle.net/10451/64777>. Acesso em: 15 abr. 2025.

RIBEIRO, Cláudio José Silva. Projeto de serviços para Gestão Documental com uso de abordagem tecnológica SOA (Arquitetura Orientada a Serviços). **Informação & Sociedade**, João Pessoa, v. 22, n. 3, 2012. Disponível em: <https://periodicos.ufpb.br/index.php/ies/article/view/10262>. Acesso em: 15 abr. 2025.

RODRIGUEZ, Guillermo *et al.* Understanding and addressing the allocation of microservices into containers: a review. **IETE Journal of Research**, [S. l.], v. 70, n. 4, p. 3887-3900, 2024.

SANTOS, Luiz Olavo Bonino da Silva; MORAIS FONSECA, Claudenir; PRINCE SALES, Tiago. HTTP-Based Implementation Strategies for the FAIR Digital Object Framework Identifier Resolution. **Open Conference Proceedings**, Berlin, v. 5, 2025. DOI: <https://doi.org/10.52825/ocp.v5i.1423>. Disponível em: <https://www.tib-op.org/ojs/index.php/ocp/article/view/1423>. Acesso em: 15 abr. 2025.

SANTOS, Luiz Olavo Bonino da Silva. **FAIR Digital Object Framework Documentation - working draft**. 2022. Disponível em: <https://fairdigitalobjectframework.org> . Acesso em: 10 abr. 2025.

SCHWARDMANN, Ulrich. Digital objects–fair digital objects: Which services are required? **Data Science Journal**, [S. l.], v. 19, p. 15-15, 2020. Disponível em: <https://doi.org/10.5334/dsj-2020-015>. Acesso em: 15 abr. 2025.

WEIGEL, Tobias.; SCHWARDMANN, Ulrich.; KLUMP, Jens. *et al.* Making Data and Workflows Findable for Machines. **Data Intelligence**, [S. l.], v. 2, n. 1-2, p. 40-46. 2020. DOI: https://doi.org/10.1162/dint_a_00026. Disponível em: https://www.sciengine.com/DI/doi/10.1162/dint_a_00026. Acesso em: 07 maio 2025.

WILKINSON, Mark D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. **Scientific Data**, [S. l.], v. 3, 2016.. Disponível em: <https://www.nature.com/articles/sdata201618>. Acesso em: 15 abr. 2025.

World Wide Web Consortium/DCAT (W3C). **RDF**. 201[?]. Disponível em: <https://www.w3.org/TR/vocab-dcat-3/>. Acesso em: 15 abr. 2025.

AGRADECIMENTOS

^A Bolsista de Iniciação Científica do CNPq/Unirio.

^B Coordenador deste estudo com apoio da FAPERJ (Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro), Processo SEI E-26/204.049/2024 e pelo CNPq por meio de Bolsa de Produtividade PQ2/CNPq.



[4.0 Internacional](https://creativecommons.org/licenses/by/4.0/)

Esta licença permite compartilhamento, remixe, adaptação e criação a partir do trabalho, mesmo para fins comerciais, desde que sejam atribuídos créditos ao(s) autor(es). Conteúdos elaborados por terceiros, citados e referenciados nesta obra não são cobertos pela licença.

Como citar este trabalho:

AMORIM, David Augusto Maduro de Mesquita de; FORAIN, Lucas dos Santos; RIBEIRO, Cláudio José Silva. Desafios da implementação de um Fair Digital Object (FDO): uma proposta de arquitetura para encapsular dados e metadados. *In*: WORKSHOP DE INFORMAÇÃO DADOS E TECNOLOGIA, 8., 2025, Marília, SP. **Anais [...]**. Marília, SP: Universidade de Marília, 2025. DOI: <http://doi.org/10.22477/viii.widat.231>.



WIDaT 2025

VIII WORKSHOP DE INFORMAÇÃO,
DADOS E TECNOLOGIA

MARÍLIA • SP

Modelo de avaliação diagnóstica FAIR para fontes de informação científicas na internet endereçadas a bibliotecas digitais abertas e agregadas

*FAIR diagnostic assessment model for scientific information sources on the internet
addressed to open and aggregated digital libraries*

Heloene Gonçalves Santos Passos¹



<https://lattes.cnpq.br/2362442830790668>



<https://orcid.org/0009-0009-5603-3323>

Daniela Lucas da Silva Lemos²



<http://lattes.cnpq.br/9280443047358807>



<https://orcid.org/0000-0003-1565-7366>

Dirceu Flávio Macedo³



<http://lattes.cnpq.br/2765166629600107>



<https://orcid.org/0000-0003-4837-3618>

Resumo

O artigo propõe um modelo de avaliação diagnóstica baseado nos princípios FAIR (*Findable, Accessible, Interoperable, Reusable*) para avaliar fontes de informação na internet, com foco na qualificação de acervos em bibliotecas digitais abertas e agregadas. A pesquisa, de natureza aplicada, adota uma abordagem quali-quantitativa, exploratória e descritiva, analisando periódicos científicos da área da Museologia. O modelo se fundamenta em seis dimensões analíticas: (i) Padrão de Dados, (ii) Metadados e Indexação, (iii) Licença de Uso, (iv) Infraestrutura e Protocolos de Acesso, (v) Documentação e Proveniência, e (vi) Integração e Conectividade. A aderência FAIR foi mensurada por uma fórmula que avalia os critérios dentro dessas dimensões, utilizando uma escala de pontuação de 0 a 3. Os resultados revelam que 20% dos periódicos analisados atingiram a aderência máxima aos princípios FAIR. Destacou-se que a adoção de procedimentos simples, como o uso de identificadores persistentes, poderia aumentar significativamente a conformidade dos demais periódicos. Além disso, periódicos que utilizam padrões como *Dublin Core* e protocolos interoperáveis apresentaram melhor indexação e acessibilidade. O estudo conclui que o modelo proposto é uma ferramenta eficiente para avaliar a qualidade de fontes de informação e apoiar a seleção, a extração e a organização de coleções em bibliotecas digitais. Os achados oferecem subsídios para gestores e pesquisadores das áreas de Ciência Aberta, Humanidades Digitais e Ciência da Informação, auxiliando na tomada de decisões e na implementação de boas práticas para o aprimoramento da interoperabilidade e reutilização de dados científicos na Web.

¹ Mestranda em Ciência da Informação, Universidade Federal do Espírito Santo (UFES), Vitória, ES, Brasil. Assistente Administrativo do Sistema Integrado de Biblioteca, Universidade Federal do Espírito Santo (UFES), Vitória, ES, Brasil. heloene.passos@ufes.br.

² Doutora em Ciência da Informação, Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, MG, Brasil. Professora e Pesquisadora, Universidade Federal do Espírito Santo (UFES), Vitória, ES, Brasil. daniela.l.silva@ufes.br.

³ Mestre em Ciência da Informação, Universidade Federal do Espírito Santo (UFES), Vitória, ES, Brasil. Pesquisador no Laboratório de Ciência da Informação, Dados e Tecnologia (Cidat Lab), Universidade Federal do Espírito Santo (UFES), Vitória, ES, Brasil. dfmacedo@gmail.com.

Palavras-chave: Princípios FAIR; análise de conteúdo; Modelo de Avaliação Diagnóstica FAIR; fontes de informação; biblioteca digital.

Abstract

The article proposes a diagnostic evaluation model based on the FAIR principles (Findable, Accessible, Interoperable, Reusable) to assess information sources on the internet, focusing on the qualification of collections in open and aggregated digital libraries. The research, applied in nature, adopts a qualitative, exploratory, and descriptive approach, analyzing scientific journals in the field of Museology. The model is based on six analytical dimensions: (i) Data Standards, (ii) Metadata and Indexing, (iii) Usage License, (iv) Infrastructure and Access Protocols, (v) Documentation and Provenance, and (vi) Integration and Connectivity. FAIR adherence was measured using a formula that evaluates criteria within these dimensions, employing a scoring scale from 0 to 3. The results reveal that 20% of the analyzed journals achieved the highest level of adherence to the FAIR principles. It was highlighted that adopting simple procedures, such as the use of persistent identifiers, could significantly improve compliance among other journals. Furthermore, journals that apply standards such as Dublin Core and interoperable protocols demonstrated better indexing and accessibility. The study concludes that the proposed model is an effective tool for evaluating the quality of information sources and supporting the selection, extraction, and organization of collections in digital libraries. The findings provide valuable insights for managers and researchers in the fields of Open Science, Digital Humanities, and Information Science, aiding decision-making and the implementation of best practices to enhance interoperability and the reuse of scientific data on the Web.

Keywords: FAIR Principles; content analysis; FAIR Diagnostic Evaluation Model; information sources; digital library.

1 INTRODUÇÃO

O grande volume de dados gerados pela explosão da informação tem sido um desafio no aspecto da organização e recuperação da informação com qualidade para instituições, governos, pesquisadores e sociedade em geral, que demonstram preocupação com o fluxo informacional presente nos meios digitais devido à percepção de que, muitas vezes, não se verifica a adoção de padrões e regras (Gilliland, 2016) na disponibilização dos conteúdos nas bases de dados.

Lemos e Souza (2024) consideram que encaminhamentos de pesquisas têm sido feitos progressivamente pela Ciência da Informação e Ciência da Computação, cujos esforços são endereçados aos sistemas de recuperação da informação (SRIs) em ambientes digitais a fim de melhorar o escopo de interoperabilidade entre diferentes esquemas de metadados e aplicações.

Nesse contexto, os princípios FAIR (Wilkinson *et al.*, 2016), um acrônimo para (*Findable*/Localizável, *Accessible*/Acessível, *Interoperable*/Interoperável e *Reusable*/Reutilizável), se adentram e podem ser considerados como um norteador de ações a serem desenvolvidas nos SRIs (Martins *et al.*, 2022), desde a seleção, a extração,

o tratamento e a organização de coleções de documentos até o depósito desses e seus registros de metadados em bases de dados, repositórios ou bibliotecas digitais, com a finalidade de garantir localização, acesso, interoperabilidade e reuso dos dados dispostos nas bases ou fontes de informação disponíveis na rede. Para Araújo e Fachin (2015), fonte de informação, vista como um sistema de informação, pode ser um documento, base de dados ou um repositório, tendo a característica de informar algo para alguém.

A disponibilização de fontes de informação na rede para uso e reúso tem forte relação com o movimento Ciência Aberta, que vai ao encontro dos objetivos da Unesco (Unesco, 2022) em valorizar a colaboração e o compartilhamento de conhecimento à humanidade, baseando-se nos sistemas de propriedade intelectual para ampliar o acesso de todos ao conhecimento para o benefício da ciência e da sociedade. A preocupação da Ciência Aberta é “tornar a atividade de pesquisa mais transparente, mais colaborativa e mais eficiente” (Sayão; Sales, 2021, p. 175). Assim sendo, os princípios FAIR facilitam a governança de dados, a cooperação internacional e o acesso universal ao conhecimento científico, pois objetiva garantir a localização, o acesso, a interoperabilidade e o reúso de artefatos digitais disponibilizados pela comunidade científica na rede ou em rede, assumindo, logo, um importante reconhecimento internacional nas boas práticas de gestão de dados em diferentes contextos (Sales *et al.*, 2021).

Adicionalmente, os avanços tecnológicos e a Web, a exemplo do FAIR, possibilitaram um maior alcance da informação de forma rápida, dinâmica e acessível. As Bibliotecas Digitais (BD) são exemplos desse alcance, disponibilizando informação além do espaço físico. Essa característica das Bibliotecas Digitais expande suas possibilidades de alcance fazendo com que se tornem um meio eficaz para distribuição, cooperação e acesso ao conhecimento científico, além de atuarem como núcleo agregador para comunidades segmentadas e distribuídas geograficamente (Sayão, 2021).

Nessa perspectiva, iniciativas de criação e disponibilização de bibliotecas digitais com dados abertos, agregados e científicos têm crescido nos últimos anos, especialmente no âmbito do patrimônio cultural digital (Martins *et al.*, 2022), proporcionando inovação na forma de produção e tratamento de fontes de informação.

Entretanto, um dos principais desafios na construção de BD com dados abertos e agregados de fontes heterogêneas encontra-se na localização e na seleção de fontes de

informações pertinentes ao propósito desse tipo de SRI, além da extração de dados contextualizados e da organização desses dados em coleções temáticas a compor a base de dados agregada (Martins *et al.*, 2022).

Tal cenário dificulta a prestação de serviço de qualidade aos usuários, o que implica em garantir que tenham acesso a informações precisas, atualizadas, acessíveis e confiáveis, contribuindo, portanto, para uma experiência positiva em relação ao SRI. Além disso, dificulta a reutilização de fontes de informação na Web, considerando que os objetos digitais são elementos preponderantes na geração de valor em suas circulações e reuso social. Isso, inclusive, prejudica o avanço do crescimento de campos emergentes, como as Humanidades Digitais (Poole, 2017), que buscam a extração de valor em objetos digitais por meio de ações de reuso para o desenvolvimento econômico, social e humano.

Nesse ensejo, formula-se a questão desta pesquisa em responder: como avaliar a qualidade de fontes de informações científicas para fins de seleção, extração e organização de coleções temáticas em bibliotecas digitais abertas e agregadas à luz dos princípios FAIR? As fontes de informações científicas escolhidas como amostragem para o percurso da presente pesquisa foram as da área da Museologia pelo fato de que há em andamento um projeto de pesquisa mais abrangente envolvendo a construção de uma biblioteca digital aberta e agregada de conhecimento científico para este domínio.

Assim sendo, o objetivo da presente pesquisa é apresentar um modelo de avaliação diagnóstica fundamentado nos princípios FAIR para a avaliação de fontes de informação científicas de interesse à criação de Bibliotecas Digitais abertas e agregadas.

Os resultados desta pesquisa podem ajudar o gestor de um projeto envolvendo BD desta natureza a visualizar, de forma sistematizada, a relação entre fontes de informações candidatas a re(uso), possíveis dimensões analíticas sobre áreas importantes que refletem a qualidade de fontes de informação científica na internet e índices de aderência aos princípios FAIR. Essa visualização pode ajudar o gestor a identificar quais áreas (dimensões) apresentam maior conformidade e desafios, segundo os princípios FAIR, facilitando a tomada de decisões e a análise comparativa entre fontes de informação candidatas a compor uma BD.

Para além, essas dimensões podem ajudar na investigação se as fontes de informação seguem os princípios FAIR de maneira eficaz. Uma abordagem sistemática

com métricas de informação pode revelar quais fontes estão mais alinhadas às boas práticas e quais precisam de melhorias para se tornarem mais encontráveis, acessíveis, interoperáveis e reutilizáveis.

2 PROCEDIMENTOS METODOLÓGICOS

A presente pesquisa é classificada como de natureza aplicada, de abordagem quali-quantitativa, de objetivo exploratório e descritivo em sites de periódicos nacionais da área da Museologia, adotando pesquisa bibliográfica, documental e mapeamento de fontes de informação.

O primeiro ponto a considerar nas etapas metodológicas foi a concepção das dimensões analíticas, para as quais se usou o método de análise de conteúdo (Bardin, 2016) em artigos científicos ligados à Ciência da Informação que retratam áreas importantes sobre qualidade de fontes de informação científicas na internet para uso e reúso, com garantia literária, conforme revisão bibliográfica. As dimensões são enumeradas e elucidadas abaixo:

- a) **Padrão de dados:** recomendações para a padronização de estrutura (ex.: padrão de metadados *Dublin Core*), valor (ex.: vocabulário controlado *Art & Architecture Thesaurus* - AAT), conteúdo (ex.: regra de catalogação *Anglo American Cataloguing Rules* - AACR2) e comunicação de dados, numa linguagem legível para a máquina (ex.: *Dublin Core Extensible Markup Language* - XML), para a produção qualificada de metadados.
- b) **Indexação a partir de metadados:** adoção de padrões de metadados; descritores preenchidos e padronizados; compatibilidade com mecanismos de busca e *harvesters*.
- c) **Licença de uso:** especificação do conjunto de licenças explicitadas para publicação dos dados (ex.: *Copyright*; *Creative Commons*); indicação clara das permissões e restrições de uso.
- d) **Infraestrutura tecnológica e de comunicação:** especificação do tipo de sistema de recuperação da informação e *softwares* usados (ex.: repositório *Open Journal Systems* - OJS; protocolos de busca e recuperação, tais como o

Hypertext Transfer Protocol – HTTP e o *Open Archives Initiative Protocol for Metadata Harvesting* - OAI-PMH; *Application Programming Interface* – APIs para acesso à base de dados) para acessar e manipular dados e metadados.

- e) **Documentação e proveniência:** disponibilidade de documentação clara sobre os dados e a fonte de informação (ex.: catálogo bibliográfico; repositório de dados de pesquisa); e registro de proveniência dos dados (ex: quem criou; processou; modificou; versões anteriores).
- f) **Integração e Conectividade:** uso de identificadores padronizados para autores, instituições e publicações (ex: *Open Researcher and Contributor ID* - ORCID; *Research Organization Registry* -ROR; *International Standard Serial Number*- ISSN; *Digital Object Identifier* - DOI); e capacidade de interligar diferentes conjuntos de dados e metadados abertos (ex.: princípios para *linked open data*).

A segunda etapa foi determinar o critério de seleção das fontes de informação para compor a lista de avaliação (*corpus* central). Assim, dentre vários tipos de fontes passíveis de serem identificadas e selecionadas na rede, o critério adotado foi a seleção inicial de 10 repositórios de periódicos científicos da área da Museologia, visto que a presente pesquisa faz parte de um projeto maior em parceria com o Instituto Brasileiro de Museus (Ibram) em criar uma biblioteca digital aberta e agregada de conhecimento científico dos museus brasileiros. Deste modo, seria possível obter resultados plausíveis para validação e avanços nos estudos.

A última etapa metodológica focou na construção do modelo de avaliação diagnóstica das fontes de informação (periódicos científicos) para o qual contou com uma planilha de dados para pontuação, sendo formatada, em suas linhas, com a relação das fontes de informação, e suas colunas com as 6 dimensões analíticas, cada qual com subníveis alinhados aos critérios FAIR, totalizando 15 variáveis para cada dimensão. É válido ressaltar que os critérios FAIR não se relacionam com apenas uma dimensão. Frequentemente, eles estão presentes em outras dimensões, tornando necessária uma análise rigorosa e criteriosa da aderência desses princípios em áreas importantes da Ciência da Informação. Essas áreas contribuem para a qualidade das fontes de informações científicas disponíveis na Internet para diversos fins, com destaque para a

extração e agregação de dados provenientes de fontes heterogêneas em bases de dados. O Quadro 1 a seguir exhibe a relação dos critérios FAIR juntamente com suas respectivas descrições.

Quadro 1 – Critérios FAIR e suas descrições

Critério	Descrição
F1	Dados devem ter um identificador único e persistente (PID – <i>persistent identifier</i> , em inglês).
F2	Dados devem ser descritos com metadados ricos e precisos.
F3	Metadados devem incluir o identificador do dado que descrevem.
F4	Metadados devem ser registrados ou indexados em um recurso pesquisável.
A1	Dados devem ser recuperáveis por meio de protocolos padronizados.
A1.1	O protocolo deve ser aberto, livre e universalmente implementável.
A1.2	O protocolo deve permitir autenticação e autorização, quando necessário.
A2	Metadados devem permanecer acessíveis, mesmo que os dados não estejam mais disponíveis.
I1	Dados devem usar linguagens, formatos e vocabulários padronizados e amplamente aceitos.
I2	Dados devem usar vocabulários que sigam os princípios FAIR.
I3	Dados devem incluir referências a outros dados para facilitar a integração.
R1	Dados devem ser descritos com atributos precisos e relevantes
R1.1	Dados devem ter uma licença clara e acessível que permita reutilização
R1.2	Dados devem incluir informações sobre sua proveniência.
R1.3	Dados devem seguir padrões de domínio relevantes para permitir reutilização.

Fonte: Elaboração própria.

A escala de pontuação adotada variou de 0 a 3, sendo: **0**, quando o critério FAIR não se aplicava à dimensão avaliada, segundo o julgamento do avaliador; **1**, para casos de baixa aderência; **2**, quando a aderência era considerada média; e **3**, nos casos de alta aderência. A análise foi conduzida de forma manual e criteriosa em fevereiro de 2025, com posterior processamento automático dos dados para cálculo da pontuação atribuída a cada dimensão da fonte de informação.

O cálculo seguiu a seguinte fórmula: Pontuação $i,k = ((\sum \text{Valor } i,j,k) / n)$

Onde:

- “i” é uma **fonte de informação** particular.
- “j” é um **critério** aplicável particular incluído na **dimensão**.
- “k” é uma **dimensão** na **fonte de informação** “i”.

- Pontuação_k é a pontuação para a fonte de informação “i” para o conjunto de critérios pertencente à dimensão “k”.
- Valor_{ij,k} é o valor para o critério “j” de uma dimensão “k” na fonte de informação “i”.
- n é o número de **critérios aplicáveis** a uma **dada dimensão**.

Após a aplicação da pontuação final (dada pela média) para os 10 periódicos científicos, foi possível obter a informação de *ranking*, o que permitiu avaliá-los frente às dimensões analíticas, apresentadas e discutidas na próxima seção.

3 RESULTADOS E DISCUSSÃO

Os resultados foram obtidos a partir da análise do avaliador em relação aos critérios FAIR aplicáveis e não aplicáveis, o que resultou no arranjo que se segue. Em Padrão de Dados, foram identificados 12 critérios aplicáveis e 3 não aplicáveis (F3, A2, R1). Em Metadados e Indexação, foram 13 critérios aplicáveis e 2 não aplicáveis (A1.2, R1.3). Na Licença de Uso, houve 9 critérios aplicáveis e 6 não aplicáveis (F2, F3, F4, A1.2, I2, R1.3). Em Infraestrutura e Protocolos de Acesso, foram 14 critérios aplicáveis e 1 não aplicável (R1.3). Em Documentação e Proveniência, registraram-se 10 critérios aplicáveis e 5 não aplicáveis (F3, F4, A1.2, R1, R1.3). Por fim, em Integração e Conectividade, houve 13 critérios aplicáveis e 2 não aplicáveis (F3, R1.3).

Assim, para os 10 periódicos analisados, dentro das 6 categorias analíticas propostas, constataram-se diferentes níveis de conformidade com os princípios FAIR, variando entre 1 e 3, conforme mostrado no Quadro 2.

Quadro 2 – Ranking dos periódicos científicos frente às dimensões baseadas em FAIR

Periódico	Pontuação	Análise dos periódicos com base nas seis dimensões analíticas e nos princípios FAIR
Anais do Museu Paulista: História e Cultura Material	3	Emprega padrão de metadados <i>Dublin Core</i> . Metadados descritivos e administrativos preenchidos de forma completa e normalizada, o que sugere uso de vocabulário controlado. Apresenta metadados indexados em mecanismos de busca. Licenças de uso claras. Utiliza o <i>software</i> OJS com suporte ao protocolo OAI-PMH. Apresenta documentação de dados e a proveniência é verificada apenas pelo uso de referências bibliográficas nos artigos. Faz uso de PID.

Boletim do Museu Paraense Emílio Goeldi Ciências Naturais	3	Emprega padrão de metadados <i>Dublin Core</i> . Metadados descritivos e administrativos preenchidos de forma completa e normalizada, o que sugere uso de vocabulário controlado. Apresenta metadados indexados em mecanismos de busca. Licenças de uso claras. Utiliza o <i>software</i> OJS com suporte ao protocolo OAI-PMH. Apresenta documentação de dados e a proveniência é verificada apenas pelo uso de referências bibliográficas nos artigos. Faz uso de PID.
Boletim do Museu Paraense Emílio Goeldi Ciências Humanas	2,8	Metadados descritivos e administrativos preenchidos de forma completa e normalizada, o que sugere uso de vocabulário controlado. Apresenta metadados indexados em mecanismos de busca. Licenças de uso claras. Apresenta documentação de dados e a proveniência é verificada apenas pelo uso de referências bibliográficas nos artigos. Faz uso de PID.
Anais do Museu Histórico Nacional	2,7	Emprega padrão de metadados <i>Dublin Core</i> . Metadados descritivos e administrativos preenchidos. Apresenta metadados indexados em mecanismos de busca. Licenças de uso claras. Apresenta documentação de dados. Usa ORCID e ISSN.
Humanidades em Revista	2,16	Emprega padrão de metadados <i>Dublin Core</i> . Metadados descritivos e administrativos preenchidos. Apresenta metadados indexados em mecanismos de busca. Licenças de uso claras. A proveniência é verificada apenas pelo uso de referências bibliográficas nos artigos. Usa apenas o ISSN.
Arquivos do Museu de História Natural e Jardim Botânico da UFMG	2,07	Emprega padrão de metadados <i>Dublin Core</i> . Metadados descritivos e administrativos preenchidos. Apresenta metadados indexados em mecanismos de busca. Utiliza o <i>software</i> OJS com suporte ao protocolo OAI-PMH. A proveniência é verificada apenas pelo uso de referências bibliográficas nos artigos. Usa apenas o ORCID.
Boletim do Museu de Biologia Mello Leitão	2,03	Usa apenas ISSN. Apresenta metadados indexados em mecanismos de busca. Licenças de uso claras. Apresenta documentação de dados. A proveniência é verificada apenas pelo uso de referências bibliográficas nos artigos.
Africanidades: a Revista do Mafró	1,99	Usa apenas ISSN. Emprega padrão de metadados <i>Dublin Core</i> . Metadados descritivos e administrativos preenchidos. Apresenta metadados indexados em mecanismos de busca. Licenças de uso claras. Utiliza o protocolo OAI-PMH. A proveniência é verificada apenas pelo uso de referências bibliográficas nos artigos.
Anuários do Museu Imperial (1940 – 2022)	1,97	Usa apenas ISSN. Apresenta metadados indexados em mecanismos de busca. A proveniência é verificada apenas pelo uso de referências bibliográficas nos artigos.
Boletim Informativo do Museu de Arqueologia e Etnologia da Universidade Federal da Bahia	1,69	Usa apenas o ISSN.

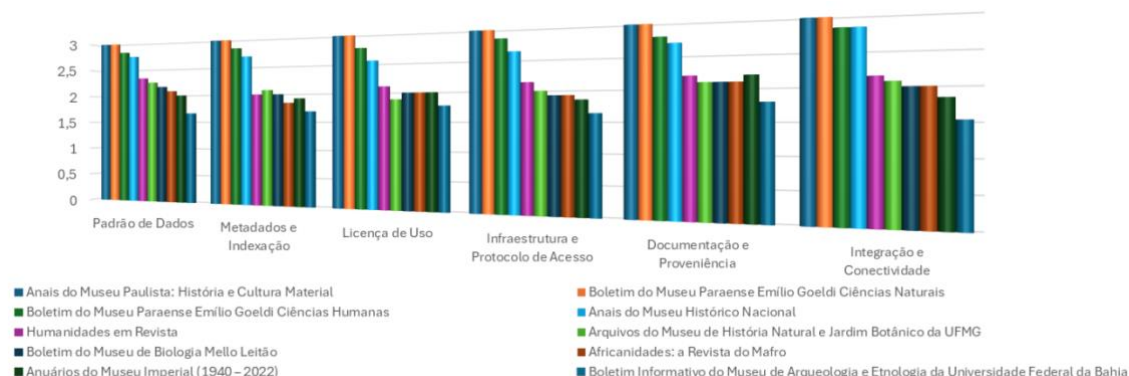
Fonte: Elaboração própria.

Todos os periódicos analisados utilizam protocolo de comunicação padronizada, aberto, gratuito e universalmente implementável (HTTP) e linguagem formal, acessível, compartilhada e amplamente aplicável (XML), viabilizando aspectos relacionados à comunicação de seus dados na rede. Destacam-se os periódicos Anais do Museu Paulista: História e Cultura Material e Boletim do Museu Paraense Emílio Goeldi Ciências Naturais,

que apresentam máxima adesão (3,0) aos princípios nas 6 categorias avaliadas. Já os periódicos Boletim do Museu Paraense Emílio Goeldi Ciências Humanas e Anais do Museu Histórico Nacional têm uma média de boa aderência também, com pontuações de 2,8 e 2,7, respectivamente. Os periódicos Humanidades em Revista, Arquivos do Museu de História Natural e Jardim Botânico da UFMG, Boletim do Museu de Biologia Mello Leitão, Africanidades: a Revista do Mafo, e Anuários do Museu Imperial (1940 – 2022) obtiveram pontuações moderadas de 2,16; 2,07; 2,03; 1,99; e 1,97, respectivamente. Finalmente, o periódico Boletim Informativo do Museu de Arqueologia e Etnologia da Universidade Federal da Bahia recebeu a avaliação mais baixa, com a pontuação de 1,69.

O Gráfico 1 sintetiza os resultados das análises, apresentando a conformidade dos periódicos científicos com as dimensões analíticas baseadas em FAIR. A análise permite identificar os periódicos que atendem melhor aos critérios de Encontrabilidade, Acessibilidade, Interoperabilidade e Reutilização, além de destacar áreas (dimensões analíticas) que necessitam de melhorias.

Gráfico 1 – Análises de cada fonte de informação frente às dimensões baseadas em FAIR



Fonte: Elaboração própria.

Em primeiro lugar, a aplicação dos princípios FAIR como critérios de avaliação permitiu distinguir os repositórios que estão mais alinhados às boas práticas dos que necessitam de melhorias. Os repositórios com pontuação máxima, como os Anais do Museu Paulista: História e Cultura Material e Boletim do Museu Paraense Emílio Goeldi Ciências Naturais, demonstraram uma forte aderência aos princípios FAIR em todas as categorias. Isso se reflete principalmente na implementação eficaz de identificadores persistentes como ORCID e DOI, bem como na utilização de licenças adequadas para garantir a acessibilidade e a reutilização dos dados (Marín-Arraiza; Heredia, 2021).

Por outro lado, repositórios com pontuação inferior, como Boletim Informativo do Museu de Arqueologia e Etnologia da Universidade Federal da Bahia e Anuários do Museu Imperial (1940 - 2022), mostraram lacunas significativas, particularmente nas categorias licença de uso, padrão de dados, e metadados e indexação. A ausência de licenças claras e a falta de metadados completos prejudicam a recuperação e a reutilização da informação, evidenciando a necessidade de atenção nesses aspectos com o objetivo de melhorar a aderência aos princípios FAIR.

A análise das métricas também destacou a importância da padronização dos dados e da infraestrutura tecnológica para a integração e conectividade dos repositórios. Os periódicos que adotam o padrão *Dublin Core*, protocolos como o OAI-PMH, e repositório digital OJS apresentam uma melhor indexação e interoperabilidade, facilitando a comunicação e a troca de dados (Gilliland, 2016) entre diferentes sistemas e repositórios. Isso é fundamental para o desenvolvimento de coleções digitais que sejam consistentes, bem-organizadas e facilmente acessíveis.

Em suma, as métricas aplicadas não apenas identificaram os repositórios que seguem as boas práticas, mas também forneceram um diagnóstico claro das áreas que precisam de melhorias. Esse diagnóstico é essencial para orientar a tomada de decisões e implementar ações corretivas que melhorem a qualidade e a eficácia dos repositórios, contribuindo para o desenvolvimento de coleções digitais mais robustas e funcionais.

Finalmente, essa abordagem permite que gestores ou profissionais da informação identifiquem quais áreas apresentam maior conformidade e desafios em relação aos princípios FAIR, facilitando a tomada de decisões e a análise comparativa entre diferentes fontes de informação endereçadas, por exemplo, a uma BD aberta e agregada.

4 CONSIDERAÇÕES FINAIS

Buscou-se, nesta pesquisa, a proposição de um modelo de avaliação diagnóstica genérico e escalável por meio do qual poder-se-ia reproduzi-lo para outros cenários que lidam com fontes de informação na internet e que demandam a apuração de suas qualidades. Considera-se, portanto, o modelo genérico uma vez que a apuração dos resultados permite avaliar quaisquer dimensões e critérios que se fizerem necessários a

um determinado contexto de pesquisa envolvendo qualidade de fontes de informação, além de levar em consideração somente os critérios FAIR que são aplicáveis a uma dimensão em particular, segundo o ponto de vista de quem está executando a avaliação; e escalável porque permite o seu avanço para outros tipos de fontes de informação, bases de dados, sistemas de informação, enfim, para além dos periódicos científicos.

É importante destacar que este estudo se restringiu a um *corpus* composto por 10 periódicos científicos. Apesar dessa limitação, o conjunto analisado revelou-se adequado e coerente com os resultados obtidos, configurando-se como um pré-teste relevante para a continuidade da pesquisa, especialmente quando se considerar a ampliação do *corpus* e a inclusão de outros tipos de fontes de informação. No que se refere à subjetividade e à dependência do julgamento do avaliador durante o processo de avaliação, recomenda-se, como perspectiva para estudos futuros, o desenvolvimento de regras explícitas para aplicação das 15 variáveis FAIR em cada uma das dimensões envolvidas no modelo de avaliação diagnóstica. Dessa forma, os procedimentos metodológicos poderão ser ainda mais alinhados aos resultados obtidos, favorecendo maior consistência e reprodutibilidade da análise.

Frente ao exposto, tanto a questão quanto o objetivo de pesquisa foram atendidos na apresentação de um modelo de avaliação diagnóstica fundamentado nos princípios FAIR para a avaliação de fontes de informação científicas para fins de seleção, extração e organização de coleções temáticas em Bibliotecas Digitais abertas e agregadas.

Espera-se que os resultados aqui apresentados possam ser úteis às comunidades de Ciência Aberta, Humanidades Digitais, Ciência da Informação e outras interessadas em usar e reusar fontes de informação com qualidade para propósitos de pesquisas diversos. Para além, os resultados serão úteis à continuidade de um estudo mais abrangente envolvendo a construção de uma Biblioteca Digital Temática ao Ibram, incluindo ações de mapeamento de fontes de informação estratégicas, especialmente documentos associados à área da Museologia, incluindo periódicos científicos, teses e dissertações, revistas especializadas, entre outras; coleta dos dados mapeados nessas fontes previamente avaliadas pelo modelo diagnóstico; e agregação de dados coletados com outros dados oriundos de outras fontes documentais.

REFERÊNCIAS

- ARAÚJO, Nelma Camêlo; FACHIN, Juliana. Evolução das fontes de informação. **Biblos Revista do Instituto de Ciências Humanas e da Informação**, [S. l.], v. 29, n. 1, 2015. Disponível em: <https://periodicos.furg.br/biblos/article/view/5463/3570>. Acesso em: 27 fev.2025.
- BARDIN, Laurence. **Análise de conteúdo**. Lisboa: Edições 70, 2016.
- GILLILAND, Anne J. Setting the Stage. In: BACA, Murta. (ed.). **Introduction to metadata**. Los Angeles: Getty Research Institute, 2016. Disponível em: <https://www.getty.edu/publications/intrometadata/setting-the-stage/>. Acesso em: 27 fev. 2025.
- LEMOS, Daniela Lucas da Silva; SOUZA, Renato Rocha. Ontologies for semantic annotation: proposal for an ontological multimedia reference model. **Knowledge Organization**, [S. l.], v. 51, n. 8, p. 561-581, 2024. DOI: <https://doi.org/10.5771/0943-7444-2024-8-561>. Disponível em: <https://www.imrpress.com/journal/KO/51/8/10.5771/0943-7444-2024-8-561>. Acesso em: 27 fev. 2025.
- MARÍN-ARRAIZA, Paloma; HEREDIA, Ana. FAIR PIDs: o papel da ORCID no fortalecimento dos Princípios FAIR. In: SALES, Luana Farias *et al.* (org.). **Princípios FAIR aplicados à gestão de dados de pesquisa**. Rio de Janeiro: Ibict, 2021. p. 15-26. Disponível em: <https://www.arca.fiocruz.br/handle/icict/51281>. Acesso em: 15 fev. 2025.
- MARTINS, Dalton Lopes *et al.* Information organization and representation in digital cultural heritage in Brazil: systematic mapping of information infrastructure in digital collections for data science applications. **Journal of the Association for Information Science and Technology**, [S. l.], 2022. DOI: <https://doi.org/10.1002/asi.24650>. Disponível em: <https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.24650>. Acesso em: 18 maio 2025.
- POOLE, Alex. H. The conceptual ecology of digital humanities. **Journal of Documentation**, [S. l.], v. 73, n. 1, 2017. Disponível em: https://www.researchgate.net/publication/311921946_The_Conceptual_Ecology_of_Digital_Humanities. Acesso em: 27 fev. 2025.
- UNESCO. **Recomendação da UNESCO sobre ciência aberta**. 2022. Disponível em: https://unesdoc.unesco.org/ark:/48223/pf0000379949_por?locale=en. Acesso em: 20 fev. 2025.
- SALES, Luana Sales *et al.* Um panorama histórico da iniciativa GO FAIR: da Europa para o Brasil. In: SALES, Luana Farias *et al.* (org.). **Princípios FAIR aplicados à gestão de dados de pesquisa**. Rio de Janeiro: Ibict, 2021. p.15-26. DOI: 10.22477/9786589167242. Disponível em: <https://www.arca.fiocruz.br/handle/icict/51281>. Acesso em: 15 fev. 2025.

SAYÃO, Luis Fernando. Afinal, o que é biblioteca digital. In: SALES, Luana Farias; VIOLA, Carla Maria Martellote. **Informação digital e suas diversas abordagens pela ótica de um cientista da informação**. Rio de Janeiro: IBICT, 2021. Disponível em:

<https://www.revistas.usp.br/revusp/article/view/13709/15527>. Acesso em: 27 fev. 2025.

SAYÃO, Luis Fernando; SALES, Luana Farias. Dados abertos de pesquisa: ampliando o conceito de acesso livre. In: SALES, Luana Farias; VIOLA, Carla Maria Martellote.

Informação digital e suas diversas abordagens pela ótica de um cientista da informação. Rio de Janeiro: IBICT, 2021. Disponível em:

<https://www.reciis.icict.fiocruz.br/index.php/reciis/article/view/611/1252>. Acesso em: 27 fev. 2025.

Wilkinson, M. D. *et al.* The FAIR guiding principles for scientific data management and stewardship. **Scientific Data**, [S. l.], v. 3, n. 1, p. 160018, 2016. Disponível em:

<https://www.nature.com/articles/sdata201618>. Acesso em: 25 fev. 2025.



[4.0 Internacional](https://creativecommons.org/licenses/by/4.0/)

Esta licença permite compartilhamento, remixe, adaptação e criação a partir do trabalho, mesmo para fins comerciais, desde que sejam atribuídos créditos ao(s) autor(es). Conteúdos elaborados por terceiros, citados e referenciados nesta obra não são cobertos pela licença.

Como citar este trabalho:

PASSOS, Heloene Gonçalves Santos; LEMOS, Daniela Lucas da Silva; MACEDO, Dirceu Flávio. Modelo de avaliação diagnóstica FAIR para fontes de informação científicas na internet endereçadas a bibliotecas digitais abertas e agregadas. In: WORKSHOP DE INFORMAÇÃO, DADOS E TECNOLOGIA, 8., Marília, SP. **Anais [...]**. Marília, SP: Universidade de Marília, 2025. DOI: <https://doi.org/10.22477/viii.widat.234>.



WIDaT 2025

VIII WORKSHOP DE INFORMAÇÃO,
DADOS E TECNOLOGIA

MARÍLIA • SP

Gestão de dados abertos de pesquisa para periódicos científicos: proposição de *checklist* de conformidade com os princípios FAIR

*Management of open research data for scientific journals: proposal of a compliance
checklist with FAIR principles*

Wesley Algarve¹



<http://lattes.cnpq.br/6841463882688482>



<https://orcid.org/0000-0003-3528-8510>

Caio Saraiva Coneglian²



<http://lattes.cnpq.br/3954065076810604>



<https://orcid.org/0000-0002-6126-9113>

Resumo

Este artigo tem como objetivo propor um checklist para avaliação de conformidade às boas práticas na gestão e compartilhamento de dados de pesquisa em artigos submetidos a revistas científicas, a partir de recomendações constantes nas diretrizes dos princípios FAIR e nas políticas editoriais de periódicos indexados no *SciELO*. A construção da checklist envolveu a coleta e análise de dados dos pontos relevantes ao compartilhamento de dados de pesquisa nas políticas editoriais de periódicos *Qualis A* (A1 à A4) indexados no *SciELO*, bem como nas diretrizes dos princípios FAIR. Dos 40 periódicos selecionados para o corpus da pesquisa, 21 possuem alguma política editorial relacionada a gestão de dados abertos de pesquisa, enquanto 19 não possui qualquer política ou sequer faz menção ao tema. Enquanto a checklist, tem por finalidade auxiliar a equipe editorial de periódicos científicos que possuam políticas voltadas a gestão de dados abertos de pesquisa. A adoção de políticas de compartilhamento de dados entre os periódicos brasileiros ainda é limitada. Em relação a checklist, este mostra-se uma ferramenta que pode auxiliar tanto a equipe editorial do periódico quanto autores, por ser uma ferramenta que reúne as boas práticas em relação a gestão de dados abertos de pesquisa no contexto da submissão de artigos científicos à periódicos.

Palavras-chave: dados abertos de pesquisa; princípios FAIR; periódicos científicos; política editorial.

Abstract

This article aims to propose a checklist for assessing compliance with best practices in research data management and sharing in articles submitted to scientific journals, based on recommendations from the FAIR principles guidelines and the editorial policies of journals indexed in *SciELO*. The development of the checklist involved collecting and analyzing data on relevant aspects of research data sharing in the editorial policies of *Qualis A* (A1 to A4) journals indexed in *SciELO*, as well as in the FAIR principles guidelines. Among the 40 journals selected for the research corpus, 21 had some editorial policy related to open research data

¹ Mestrando em Ciência da Informação, Universidade Estadual Paulista Júlio de Mesquita Filho (UNESP), Marília, São Paulo, Brasil. wesley.algarve@unesp.br.

² Doutor em Ciência da Informação, Universidade Estadual Paulista Júlio de Mesquita Filho (UNESP), Marília, São Paulo, Brasil. Professor Colaborador do Programa de Pós-graduação em Ciência da Informação, Universidade Estadual Paulista Júlio de Mesquita Filho (Unesp), Marília, São Paulo, Brasil. Coordenador dos cursos de graduação de Análise e Desenvolvimento de Sistemas e Ciência da Computação da Universidade de Marília (Unimar), Marília, São Paulo, Brasil. caioconegli@unimar.br.

management, while 19 had no policy or did not mention the topic at all. The checklist is intended to assist the editorial team of scientific journals that have policies focused on open research data management. The adoption of data-sharing policies among Brazilian journals is still limited. Regarding the checklist, it proves to be a tool that can assist both journal editorial teams and authors, as it compiles best practices for open research data management in the context of article submissions to journals.

Keywords: open research data; FAIR principles; scientific journals; editorial policy.

1 INTRODUÇÃO

A ciência e a prática científica estão em contínua transformação, buscando constantemente evoluir de forma a ampliar seu impacto tanto nas comunidades científicas quanto na sociedade em geral. Uma das transformações que vem sendo experimentada nos últimos anos está relacionada a um modo mais aberto de produção e comunicação da ciência.

Esse movimento, denominado de Ciência Aberta, torna o processo do fazer científico mais democrático, eficiente e transparente, uma vez que promove o compartilhamento e acesso dos dados, metodologias e resultados das pesquisas.

Podemos definir a Ciência Aberta como “[...] um movimento colaborativo e aberto, com foco no uso da tecnologia para o compartilhamento e acesso à pesquisa” (Nascimento; Albagli, 2019, p. 7).

A Ciência Aberta possui diversas iniciativas, todas pautadas no compartilhamento e acesso ao conhecimento científico, sendo alguns exemplos dessas iniciativas o Acesso Aberto, a Pesquisa Reprodutível Aberta, a Avaliação por Pares Aberta, as Políticas de Ciência Aberta e, tema de discussão deste artigo, os Dados Abertos de Pesquisa.

Sayão e Sales (2019) indicam que, no contexto da ciência contemporânea, os dados de pesquisa deixaram de ser apenas subprodutos das pesquisas científicas para assumirem um papel central na construção de novos conhecimentos. Dessa forma, é essencial que estejam disponíveis em plataformas abertas para garantir seu reuso e preservação.

A proposta dos Dados Abertos de Pesquisa é a de justamente disponibilizar dados de pesquisas de forma a permitir que sejam livremente utilizados, reutilizados, mantidos e compartilhados, mediante reconhecimento. Esses dados devem ser

acompanhados de metadados que os descrevam, possibilitando o entendimento sobre os dados, seu contexto e autoria (UNESCO, 2021).

Esses movimentos que visam aperfeiçoar o fazer científico através do compartilhamento e a abertura dos dados de pesquisa, na prática, apenas surtirão efeito se houver boas práticas na preparação, armazenamento e a preservação dos dados. Neste sentido, surgiram os Princípios FAIR (*Findability, Accessibility, Interoperability e Reuse*), princípios que garantem que esses dados sejam encontráveis, acessíveis, interoperáveis e reusáveis.

Os princípios FAIR têm por objetivo estabelecer um conjunto de metadados que possibilite o uso tanto para sistemas computacionais automatizados quanto por usuários humanos, promovendo assim a interoperabilidade entre diferentes ambientes de dados (Henning *et al.*, 2019).

Neste contexto de boas práticas, os periódicos científicos emergem com um importante papel no incentivo à gestão de dados abertos de pesquisa, promovendo práticas transparentes na ciência e comunicação científica (Hrynaskiewicz *et al.*, 2017).

À medida que o movimento da ciência aberta ganha relevância, alguns periódicos têm atualizado suas políticas editoriais de maneira que incentivam ou mesmo exigem dos autores o compartilhamento de dados de pesquisa em repositórios, como condição para a submissão de artigos.

As agências de fomento, como a Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) no Brasil, ou internacionais como a *National Science Foundation* (NSF), o *National Institutes of Health* (NIH), têm exigido cada vez mais que os dados de pesquisas sejam compartilhados de forma a serem acessíveis e reusáveis. No caso da FAPESP, para determinadas modalidades e chamadas é necessário elaborar e enviar junto a proposta de submissão um “Plano de Gestão de Dados” (FAPESP, 2017).

Contudo, é fundamental destacar que nem todos os dados podem ser abertos e disponibilizados livremente (Muriel-Torrado; Moura, 2020). Dados sensíveis ou protegidos por direitos autorais podem conter embargos ou inviabilizar a divulgação completa. Portanto, é importante seguir o princípio “[...] o mais aberto quanto possível, o mais fechado quanto necessário” (European Commission, 2016, p. 4).

A ausência de padronização nos processos de tratamento, armazenamento, preservação e recuperação de dados também representam desafios na gestão de dados abertos de pesquisa. A adoção de práticas homogêneas e convergentes na descrição e licenciamento dos dados poderia facilitar a interoperabilidade entre instituições, promovendo uma integração transparente e reduzindo esforços adicionais (Henning *et al.*, 2019).

Nesse contexto, tornam-se questões relevantes para investigação a existência de políticas editoriais relacionadas a gestão e compartilhamento de dados em periódicos, bem como a disponibilidade de ferramentas que auxiliem na verificação de sua efetiva aplicação. No entanto, a literatura sobre publicações brasileiras nessas temáticas ainda é limitada. Essa lacuna contrasta com a crescente importância da temática no debate sobre as tendências da política de dados abertos no Brasil, evidenciando a necessidade de estudos que explorem mais profundamente essa temática.

Visando contribuir para suprir essa lacuna, este artigo tem como objetivo desenvolver uma checklist prática, fundamentada nos dados coletados das políticas editoriais de periódicos brasileiros e alinhada aos princípios FAIR. Essa ferramenta tem como finalidade auxiliar os periódicos na avaliação de artigos submetidos, verificando se estes estão aderindo às boas práticas no que se refere ao compartilhamento de dados de pesquisa. Para atingir esse objetivo, foram realizados os procedimentos metodológicos apresentados na seção a seguir.

2 PROCEDIMENTOS METODOLÓGICOS

O estudo foi realizado a partir de pesquisa documental, com abordagem qualitativa. O levantamento se deu a partir de pesquisas das políticas editoriais disponíveis nos sites dos periódicos e na página dos periódicos no *SciELO* sobre a temática de “gestão e compartilhamento de dados de pesquisa”. Os dados foram coletados no período de setembro e outubro de 2024.

No momento da coleta de dados, havia 322 periódicos indexados no *SciELO*. Para identificar o universo, utilizou-se a classificação de periódicos *Qualis* Periódicos Capes

com nota *Qualis A* (A1 à A4) do quadriênio 2017-2020, cujo total resultou em 226 periódicos com *Qualis A* indexados no *SciELO*.

O *corpus* de análise é composto por 40 periódicos brasileiros com classificação *Qualis A*, indexados no *SciELO*. Foram selecionados aleatoriamente 5 periódicos de cada uma das 8 grandes áreas do conhecimento adotadas pelo *SciELO*.

De forma a padronizar o idioma, escolheu-se consultar as páginas em inglês, uma vez que nem todas as revistas possuem a versão de suas políticas em português.

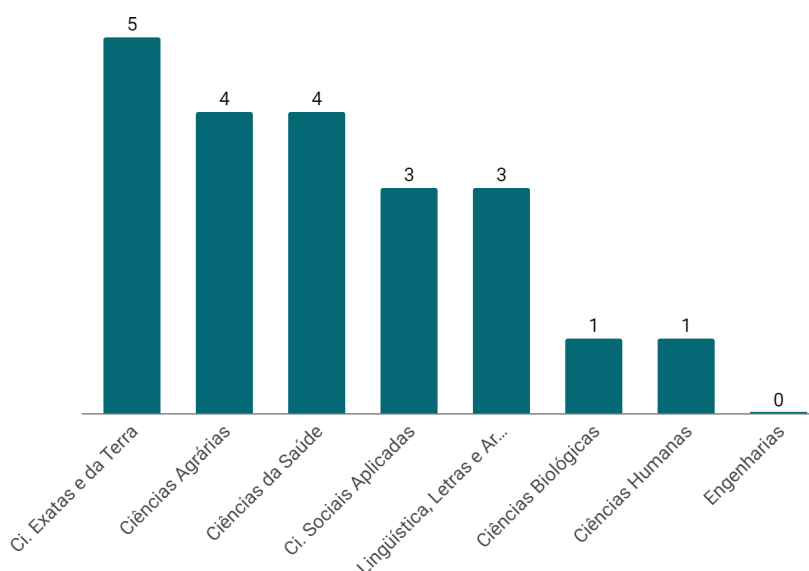
A coleta e o tratamento dos dados, realizados com o auxílio do aplicativo online Planilhas *Google*, permitiram identificar a existência de políticas relacionadas ao compartilhamento de dados de pesquisa.

A partir da análise dos dados coletados, e dos Princípios FAIR, elaborou-se a *checklist*.

3 RESULTADOS

Dos 40 periódicos selecionados para o *corpus* dessa pesquisa, 21 possuem alguma política editorial relacionada a gestão de dados abertos de pesquisa, enquanto 19 não possuem qualquer política ou sequer fazem menção ao tema.

Gráfico 1 – Periódicos que possuem políticas gestão de dados abertos de pesquisa por área



Fonte: Elaboração própria.

diretamente relacionadas aos princípios desse movimento. Termos como “*open*” (35 ocorrências) e “*share*” (25 ocorrências) reforçam a ênfase na abertura e na colaboração científica. A presença de “*repository*” (21 ocorrências) e “*deposit*” (15 ocorrências) indica uma atenção significativa à prática de arquivamento e disponibilização de dados em repositórios, processos essenciais para a reprodutibilidade e reutilização de dados na pesquisa científica. Além disso, palavras como “*transparency*” (11 ocorrências) e “*access*” (10 ocorrências) apontam para princípios norteadores da ciência aberta, como a clareza nos processos e a democratização do conhecimento.

3.1 Proposta de Instrumento para Verificação de Artigos Submetidos a Periódicos

Como observado na literatura e nas políticas, existem recomendações para a gestão do compartilhamento de dados. No entanto, tanto na literatura nacional quanto internacional, não foi identificado um instrumento que reúna essas recomendações e que auxilie os periódicos científicos. Assim, considera-se que a *checklist* atua como uma ferramenta eficaz para a verificação de artigos submetidos em periódicos, podendo ser utilizada pelas equipes editoriais em qualquer fase do processo de submissão e avaliação de artigos.

O Quadro 1 apresenta o instrumento proposto. Destaca-se as possibilidades de resposta aos critérios da *checklist*: Sim (atende ao critério); Não (não atende ao critério); e N/A (o critério não é aplicável ao artigo).

A ferramenta proposta tem como finalidade auxiliar a equipe editorial de periódicos científicos que adotam políticas voltadas à gestão de dados de pesquisa abertos. No entanto, também pode ser disponibilizada aos autores, para que estes a preencham e a enviem como documento suplementar durante a submissão de seus artigos.

A checklist foi dividida em 4 seções, sendo essas os princípios FAIR (*Findable, Accessible, Interoperable e Reusable*). Buscou-se abranger apenas os itens essenciais para assegurar os critérios de qualidade dos dados de pesquisa no contexto dos periódicos, de forma a não tornar o processo de submissão e avaliação dos artigos mais burocratizado.

A agilidade no processo editorial é fundamental para que os artigos sejam disponibilizados à comunidade de maneira mais breve possível.

Quadro 1 – Checklist de gestão de dados abertos de pesquisa para periódicos

Seção e Tópico	Item #	Sim	Não	N/A	Observações	Referência
F – Findable (Localizável)						
Os dados possuem um identificador persistente (ex.: DOI) atribuído para facilitar a <u>encontrabilidade</u> ?	F1				checklist de conformidade com os princípios FAIR	Wilkinson <i>et al.</i> (2016)
Os dados estão acompanhados de metadados ricos, contendo informações como título, autores, instituição, data de coleta, resumo e palavras-chave?	F2					Wilkinson <i>et al.</i> (2016) SciELO (2023)
O manuscrito submetido inclui clara e explicitamente referência(s) aos dados (ex.: links diretos ou citações com DOI) e vice-versa?	F3					Wilkinson <i>et al.</i> (2016)
Os dados foram depositados em repositórios indexados e acessíveis a mecanismos de busca?	F4					Wilkinson <i>et al.</i> (2016)
A – Accessible (Acessível)						
Os metadados estão acessíveis gratuitamente para download e consulta?	A1					Silveira (2022)
Os metadados estão depositados em repositórios que garantem a preservação e a acessibilidade a estes mesmo que os dados não estejam mais acessíveis?	A2					<u>FAIRsFAIR</u> (2021)
Se houverem restrições de acesso aos dados, consta uma declaração na <u>Cover Letter</u> explicando por que os dados da pesquisa não podem ser compartilhados?	A3					Políticas das revistas pesquisadas
I – Interoperable (Interoperável)						
Os dados estão disponíveis em formatos de dados abertos, com padrões internacionais documentados e amplamente utilizados (ex.: CSV, TXT, PDF, JSON, XML)?	I1					Wilkinson <i>et al.</i> (2016) SciELO (2023)
Foram utilizados vocabulários controlados para a descrição dos metadados (ex.: Dublin Core, <u>DataCite</u>)?	I2					<u>FAIRsFAIR</u> (2021)
R – Reusable (Reutilizável)						
Há descrições claras sobre as condições de reuso aos dados, tais como a licença e restrições (caso existam)?	R1					Wilkinson <i>et al.</i> (2016)
Os metadados incluem informações sobre a proveniência dos dados, como descrições sobre geração, coleta e histórico de atualizações dos dados?	R2					Henning <i>et al.</i> (2019)

Fonte: Elaboração própria.

Para a seção *Findable* (Localizável), foram selecionados 4 tópicos (F1 à F4) baseados diretamente nos princípios FAIR, presentes no artigo de Wilkinson *et al.* (2016), juntamente com uma recomendação do SciELO (2023). Buscou-se ressaltar a encontrabilidade dos dados, através da disponibilização em um repositório de dados de acesso aberto, indexado e acessível a mecanismos de busca, através da atribuição de um identificador persistente e, por fim, através da representação dos dados com metadados ricos e representativos.

A seção *Accessible* (Acessível) conta com um tópico (A1) baseado na tese de Silveira (2022), um tópico (A2) baseado nos princípios FAIR, adaptado da ferramenta FAIRsFAIR (2021), e um tópico (A3) criado após a análise das políticas editoriais das revistas pesquisadas nesta pesquisa. Devido a possíveis restrições no compartilhamento dos dados (por exemplo dados sensíveis ou protegidos por direitos autorais), esta seção estabelece diretrizes para tornar os dados acessíveis sempre que possível, garantindo apenas obrigatória a disponibilização dos metadados. Caso o compartilhamento dos dados não seja viável, é necessário que o autor apresente uma justificativa ao editor, garantindo dessa forma que os dados sejam acessíveis sempre que possíveis.

A terceira seção é a *Interoperable* (Interoperável), possui dois tópicos baseados nos princípios FAIR, sendo um (I1) de Wilkinson *et al.* (2016) e SciELO (2023) e outro (I2) em FAIRsFAIR (2021). Esta seção rege sobre o formato dos dados e a descrição dos metadados, possibilitando assim que os dados disponibilizados sejam facilmente lidos e interpretados, de forma a permitir a interoperabilidade entre diferentes conjuntos de dados.

Por fim, a seção *Reusable* (Reutilizável) contém dois tópicos (R1 e R2) baseados nos princípios FAIR, adaptados dos artigos de Wilkinson *et al.* (2016) e de Henning *et al.* (2019). De forma a incentivar e permitir a reutilização dos dados, essa seção discorre sobre as condições de reuso, como a licença e eventuais restrições, e indica para que os metadados incluam informações sobre a proveniência dos dados, como por exemplo indicações sobre geração, coleta e histórico de atualizações dos dados.

Apresentados e discutidos os resultados e a ferramenta proposta para avaliação de conformidade às boas práticas na gestão e compartilhamento de dados de pesquisa

em artigos submetidos a revistas científicas, a próxima seção traça as considerações finais a partir da pesquisa realizada.

4 CONSIDERAÇÕES FINAIS

Este estudo apresenta e analisa os resultados de uma pesquisa transversal, com abordagem qualitativa e descritiva, focada nas políticas editoriais relacionadas à gestão de dados abertos de pesquisa. A investigação foi conduzida com base em uma amostra composta por 40 periódicos brasileiros indexados na base de dados *SciELO*. O objetivo principal foi, a partir dos dados coletados das políticas editoriais, e baseado nos princípios FAIR, elaborar uma *checklist* prática que permita aos periódicos avaliarem se os artigos submetidos estão seguindo as boas práticas no que se refere ao compartilhamento dos dados de pesquisa.

Os resultados do estudo indicam que a adoção de políticas de compartilhamento de dados entre os periódicos consultados nessa pesquisa ainda é limitada. Dentre os periódicos pesquisados que possuem tais políticas, estas frequentemente são pouco desenvolvidas, carecem de estrutura e padronização, e, em muitos casos, têm caráter facultativo.

Em relação a *checklist*, esta mostra-se uma ferramenta que pode auxiliar tanto a equipe editorial do periódico quanto autores, por ser uma ferramenta que reúne as boas práticas em relação a gestão de dados abertos de pesquisa no contexto da submissão de artigos científicos à periódicos.

Como limitação do estudo, destaca-se a pequena amostra de políticas analisadas (40 periódicos), o que pode limitar a generalização dos resultados para o universo mais amplo de periódicos brasileiros. Além disso, a *checklist* proposta não foi previamente validada por profissionais que atuam diretamente na gestão editorial de periódicos. Como estudos futuros recomenda-se ampliar a amostra de periódicos e incorporar uma validação da ferramenta por profissionais da área, tais como editores e equipe editorial, aprimorando assim a efetividade e a aplicabilidade da *checklist*.

REFERÊNCIAS

- EUROPEN COMISSION. **Guidelines on FAIR data management in Horizon 2020**. [Bruxelas]: European Commission, 2016. Disponível em: https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf. Acesso em: 20 out. 2024.
- FAIRsFAIR. **Fostering FAIR data practices in Europe**. [S. l.]: DANS, 2021. Disponível em: <https://fairaware.dans.knaw.nl/>. Acesso em: 19 out. 2024.
- FUNDAÇÃO DE AMPARO À PESQUISA DO ESTADO DE SÃO PAULO (FAPESP). **Gestão de dados**. São Paulo: FAPESP, 2017. Disponível em: <http://www.fapesp.br/gestaodedados>. Acesso em: 19 out. 2024.
- HENNING, P. C. *et al.* GO FAIR e os princípios FAIR: o que representam para a expansão dos dados de pesquisa no âmbito da Ciência Aberta. **Em Questão**, Porto Alegre, v. 25, n. 2, p. 389-412, 2019. DOI: <http://doi.org/10.19132/1808-5245252.389-412>. Disponível em: <https://seer.ufrgs.br/index.php/EmQuestao/article/view/84753>. Acesso em: 23 out. 2024.
- HRYNASZKIEWICZ, I. *et al.* Standardising and harmonising research data policy in scholarly publishing. **International Journal of Digital Curation**, Edimburgo, v. 12, n. 1, p. 65-71, 2017. DOI: <http://doi.org/10.2218/ijdc.v12i1.531>. Disponível em: <http://www.ijdc.net/article/view/12.1.65>. Acesso em: 13 out 2024.
- IRAMUTEQ. **Interface de R pour les analyses multidimensionnelles de textes et de questionnaires**. IRAMUTEQ, c2025. Disponível em: <http://www.iramuteq.org/>. Acesso em: 3 maio 2025.
- MURIEL-TORRADO, E.; MOURA, R. F. Direitos Autorais e os dados de pesquisa abertos. *In*: SHINTAKU, M.; SALES, L. F.; COSTA, M. (org.). **Tópicos sobre dados abertos para editores científicos**. Botucatu, SP: ABEC, 2020. p. 137-143. DOI: <http://doi.org/10.21452/978-85-93910-04-3.cap11>. Disponível em: https://www.abecbrasil.org.br/arquivos/Topicos_dados_abertos_editores_cientificos.pdf. Acesso em: 17 out. 2024.
- NASCIMENTO, A. G.; ALBAGLI, S. Conceitos de Ciência Aberta no Brasil: uma revisão sistemática de literatura. *In*: ENCONTRO NACIONAL DE PESQUISA E PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO, 20., 2019, Florianópolis. **Anais [...]**. Florianópolis: UFSC, 2019. Disponível em: <https://conferencias.ufsc.br/index.php/enancib/2019/paper/view/1125/940>. Acesso em: 19 out 2024.
- SALES, L. F.; SAYÃO, L. F. Uma proposta de taxonomia para dados de pesquisa. **Revista Conhecimento em Ação**, [s. l.], v. 4, n. 1, p. 31-48, 2019. DOI:

<http://doi.org/10.47681/rca.v4i1.26337>. Disponível em:
<https://revistas.ufrj.br/index.php/rca/article/view/26337>. Acesso em: 26 out. 2024.

SCIELO. **Guia de preparação de dados de pesquisa**. [S. l.]: SciELO, 2023. Disponível em:
https://wp.SciELO.org/wp-content/uploads/Guia_preparacao_pt.pdf. Acesso em: 25 out. 2024.

SILVEIRA, L. **Políticas editoriais de periódicos no ecossistema da Ciência Aberta: impactos da avaliação por pares aberta, pré-impressão e dados abertos**. 2022. Projeto de teste (Doutorado em Comunicação) – Programa de Pós-Graduação em Comunicação, Faculdade de Biblioteconomia e Comunicação, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2022.

UNESCO. **Recommendation on Open Science**. Paris: UNESCO, 2021. Disponível em:
<https://unesdoc.unesco.org/ark:/48223/pf0000379949.locale=en>. Acesso em: 21 out 2024.

WILKINSON, M. D. *et al.* The FAIR guiding principles for scientific data management and stewardship. **Scientific Data**, [s. l.], v. 3, n. 60018, 2016. DOI 10.1038/sdata.2016.18. Disponível em: <https://www.nature.com/articles/sdata201618>. Acesso em: 26 out. 2024.

AGRADECIMENTOS

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.



[4.0 Internacional](https://creativecommons.org/licenses/by/4.0/)

Esta licença permite compartilhamento, remixe, adaptação e criação a partir do trabalho, mesmo para fins comerciais, desde que sejam atribuídos créditos ao(s) autor(es). Conteúdos elaborados por terceiros, citados e referenciados nesta obra não são cobertos pela licença.

Como citar este trabalho:

ALGARVE, Wesley; CONEGLIAN, Caio Saraiva. Gestão de dados abertos de pesquisa para periódicos científicos: proposição de checklist de conformidade com os princípios FAIR. *In: WORKSHOP DE INFORMAÇÃO DADOS E TECNOLOGIA*, 8., 2025, Marília, SP. **Anais [...]**. Marília, SP: Universidade de Marília, 2025. DOI: <https://doi.org/10.22477/viii.widat.264>.



WIDaT 2025

VIII WORKSHOP DE INFORMAÇÃO,
DADOS E TECNOLOGIA

MARÍLIA • SP

Arquivologia e inteligência artificial: análise das publicações e tendências no período 2020-2024

Artificial intelligence in archival science publications: trends and impacts study over a quadrennium (2020-2024)

Gabriela de Grande Guerreiro¹

 <http://lattes.cnpq.br/1514879451252079>
 <https://orcid.org/0009-0000-9844-1725>

Ana Maria dos Santos Latgé²

 <https://lattes.cnpq.br/0101249891164073>
 <https://orcid.org/0009-0003-0510-4677>

Tânia Barbosa Salles Gava³

 <http://lattes.cnpq.br/0822963734544353>
 <https://orcid.org/0000-0002-3920-5623>

Diana Vilas Boas Souto Aleixo⁴

 <http://lattes.cnpq.br/5082397017911269>
 <https://orcid.org/0000-0003-4156-354X>

Resumo

Este estudo explora a presença e o tratamento da Inteligência Artificial (IA) como ferramenta prática em publicações da área de Arquivologia, no período de 2020 a 2024. A pesquisa foi conduzida a partir da análise de textos de revistas acadêmicas de destaque no Brasil, utilizando métodos bibliométricos e análise de conteúdo para examinar a abordagem da IA nesses trabalhos. Os resultados indicam que, embora o uso da IA na gestão documental esteja em crescimento, o volume de publicações sobre o tema ainda é reduzido, especialmente no contexto brasileiro, evidenciando uma deficiência relevante na literatura da área.

Palavras-chave: inteligência artificial; arquivologia; publicações em ciência da informação.

¹ Mestra em Ciências Jurídico-Políticas, Faculdade de Direito da Universidade do Porto (FDUP), Porto, Porto, Portugal. Graduanda em Arquivologia, Universidade Federal do Espírito Santo (UFES), Vitória, ES, Brasil. gabriela.guerreiro@edu.ufes.br.

² Especialista em Cinema, Centro Universitário Belas Artes (FEBASP), São Paulo, SP, Brasil. Graduanda em Arquivologia, Universidade Federal do Espírito Santo (UFES), Vitória, ES, Brasil. analatge@gmail.com.

³ Doutora em Engenharia Elétrica - Automação, Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal do Espírito Santo (PPGEE/UFES), Vitória, ES, Brasil. Docente, Departamento de Arquivologia, Universidade Federal do Espírito Santo (CCJE), Vitória, ES, Brasil. tanias.gava@ufes.br.

⁴ Doutora em Ciência da Informação, Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual Paulista (PPGCI/UNESP), Marília, SP, Brasil. Docente, Departamento de Arquivologia, Universidade Federal do Espírito Santo (CCJE), Vitória, ES, Brasil. diana.aleixo@ufes.br.

Abstract

This study explores the presence and treatment of Artificial Intelligence (AI) as a practical tool in publications in the field of Archivology, from 2020 to 2024. The research was conducted by analyzing texts from leading academic journals in Brazil, using bibliometric methods and content analysis to examine the approach to AI in these works. The results indicate that, although the use of AI in document management is growing, the volume of publications on the subject is still small, especially in the Brazilian context, highlighting a significant deficiency in the literature in the area.

Keywords: artificial intelligence; archival science; publications in information science.

1 INTRODUÇÃO

As Tecnologias de Informação e Comunicação (TIC) têm se integrado cada vez mais à sociedade, promovendo transformações significativas em diversas áreas do conhecimento. Para Kurzweil (2015, p. 151) “[...] a inteligência artificial está à nossa volta – não temos mais a mão na tomada.” Dessa interação, emergem tecnologias mais inteligentes, como destaca Pierre Lévy (1999, p. 75) “Quando as interações podem enriquecer ou modificar o modelo, o mundo virtual torna-se um vetor de inteligência e criação coletivas.” Em um cenário global interconectado, os dados gerados são continuamente coletados, explorados e compartilhados, tornando-se insumos valiosos para que tecnologias baseadas em inteligência artificial operem de forma autônoma, conforme Weiss (2019).

Na Arquivologia, as TIC possibilitam um momento de mudança de paradigma. Taylor (1988, p. 23) já indicava tal transformação ao afirmar que *“As expert systems, the artificial intelligence, and other front-end user programmes render the complexity and sophistication of the software more ‘transparent,’ the barriers between users’ needs and their fulfilment will grow less [...]”*

Nesse contexto, este trabalho teve como objetivo investigar como a temática da Inteligência Artificial tem sido abordada nas publicações da área de Arquivologia no Brasil, no período de 2020 a 2024, com base na análise de artigos publicados em revistas científicas vinculadas ao campo. O interesse centra-se no uso IA como ferramenta na gestão de informação, na automatização de processos, no aprimoramento de pesquisa e acesso a informações, na análise de grandes volumes de dados para identificar padrões e tendências e, especialmente, às práticas voltadas à preservação digital no âmbito arquivístico.

2 PROCEDIMENTOS METODOLÓGICOS

A presente pesquisa adota procedimentos da bibliometria e combina métodos quantitativos e qualitativos para explorar a aplicação da IA em publicações do universo da Arquivologia. O foco central é verificar se, nas publicações recentes da área, há exemplos práticos de aplicação da IA.

A metodologia foi estruturada em 5 (cinco) etapas. Na Etapa 1, procedeu-se à seleção das publicações. A definição dos periódicos foi realizada por meio de uma busca direta no Google utilizando os termos “revistas+arquivologia”. A pesquisa foi conduzida a determinação dos títulos para pesquisa foi estabelecida a partir de busca direta no Google por “revistas+arquivologia”. A pesquisa foi conduzida em língua portuguesa, considerando o escopo nacional da investigação, voltado à produção científica brasileira. Nas primeiras posições, a busca inicial retornou, os sites da Associação dos Arquivistas do Estado do Rio de Janeiro (AAERJ), do curso de Arquivologia da Universidade Federal de Santa Maria (UFSM) e o blog Arquivista.blog.br, do arquivista Gustavo Lopes. Além deles, aparece também o site da revista *Ágora*. As publicações elencadas para análise foram compiladas dos três primeiros sites, descartado o site direto da revista *Ágora*. À lista final foi incluída a revista “Informações Arquivísticas”, por ser uma publicação da própria AAERJ. Dos sites de referência foram extraídos os nomes das revistas para a análise da Etapa 2.

Na Etapa 2, a pesquisa foi realizada nos mecanismos de busca internos dos periódicos selecionados, utilizando os termos “inteligência artificial” e “arquivologia”, bem como suas equivalências em espanhol, inglês e francês. O recorte temporal foi de 1º de janeiro de 2020 a 31 de março de 2024. Foram definidos os seguintes critérios de busca: 1) **Período da coleta**: 1/1/2020 – 31/3/2024 (a data mais recente disponível); 2) **Padrão de pesquisa**: listagem estabelecida; 3) **Fontes**: periódicos selecionados na etapa 1; 4) **Ordem de busca**: conforme a relevância indicada pelas próprias plataformas de busca; 5) **Local de busca**: mecanismos de busca internos de cada periódico; 6) **Palavras-chaves**: arquivologia “inteligencia artificial” (PT) “*inteligencia artificial*” (ES) *archive "artificial intelligence"* (IN) *archivistique "intelligence artificielle"* (FR).

As práticas da busca: 1) **leitura** do resumo, das palavras-chaves e da metodologia; 2) **busca no texto pelos termos** “inteligência artificial” e “arquivologia”; 3)

leitura dos parágrafos identificados, verificando nos mesmos: a) menção à IA; b) IA como tema central; c) IA relacionada a CI; d) menção específica à Arquivologia, e e) IA mencionada em contexto de aplicação prática ou não.

Três deliberações metodológicas foram adotadas nesta etapa. A primeira refere-se ao fato que foram analisadas todas as entradas apresentadas nos buscadores, independente do gênero textual. Por isso, optou-se por utilizar o termo “texto”, em vez de “artigo”, já que também foram encontrados editoriais, relatos de experiência, aulas inaugurais, entre outros. Na segunda foram excluídos os textos que não estavam disponíveis em acesso aberto. E na terceira, realizaram-se buscas diferenciadas em três periódicos, em razão de limitações técnicas. Na Revista *Comma* o sistema de busca não permite a combinação de termos. Foram localizadas 21 entradas para “archive”, mas nenhuma para “artificial intelligence”. Assim, a revista foi considerada sem resultados satisfatórios. Nas Revistas *Archivamos* e *Tábula*: os mecanismos de busca apresentaram falhas. A pesquisa foi realizada manualmente nos índices de cada edição publicada entre 2020 e 2024.

Ao final da Etapa 2, a pesquisa resultou em 56 entradas, que foram sistematizadas em uma planilha intitulada "Entradas – Geral", com os seguintes campos de análise: 1) revista, 2) nacionalidade da revista, 3) termo de busca, 4) título do texto, 5) trecho(s) do texto, 6) síntese, 7) IA tema central? (s/n), 8) referência arquivologia? (s/n), 9) uso prático / visão geral, 10) motivo da classificação, 11) metodologia texto, 12) resumo / abstract, 13) palavras-chaves e 14) link do texto.

Na etapa 3, houve a filtragem nos textos recuperados, sendo analisados os textos recuperados conforme o objetivo da pesquisa, apenas materiais de livre acesso foram incluídos. Os 56 textos foram reduzidos a 23, e sobre estes, procedeu-se à breve análise qualitativa, permitindo não apenas identificar a presença da IA nas publicações, mas também entender como essa tecnologia está sendo aplicada na prática.

Na etapa 4 os materiais coletados foram organizados dois grupos: Grupo 1 como “ferramenta prática”, e Grupo 2 IA como “conceito teórico”. E por fim, na etapa 5 realizou-se uma análise detalhada dos textos que abordam a IA como ferramenta prática, os quais passaram a compor a planilha ("IA Prática"), que inclui os campos: 1) ano texto;

2) nº citações (texto); 3) nome autor(es); 4) índice H autor(es)⁵ ; 5) formação autor(es) (graduação/ mestrado/ doutorado); 6) país autor(es); 7) instituição autor(es), e 8) contextos de adoção da IA.

3 RESULTADOS

A questão inicial sobre a análise é a própria determinação do material de base, isto é, a seleção das revistas. Ao usar os termos “revista” e “arquivologia” em português, na busca inicial, obteve-se como resultado páginas em português e, provavelmente por isso, a maioria das revistas referenciadas são publicações brasileiras. Há coerência nessa escolha por parte dos *sítes* de referência, já que eles se destinam aos pesquisadores, estudiosos e usuários do país. Não obstante, essa preferência faz os dados da pesquisa refletirem mais o cenário das publicações locais, não exaurindo o panorama global - ainda que tragam, na lista, alguns dos títulos internacionais mais conhecidos.

Foram consolidadas 44 revistas, sendo 33 nacionais e 11 estrangeiras. Do total, 2 não foram localizadas (uma brasileira, uma espanhola); 6 (brasileiras) não tiveram edição no período pesquisado, e 2 não possuem consulta aberta das publicações anteriores (um canadense, um multinacional), somando uma redução de 10 títulos. Portanto, permaneceram 34 revistas para análise (26 nacionais, 8 estrangeiras) e um total de 5660 textos publicados entre elas, distribuídos conforme o Quadro 1.

Quadro 1 - Revistas analisadas e volume de publicação

Nacionalidade	Qtd. Revistas	Nº Textos
Brasil	26	4134
Canadá	1	83
Espanha	2	273
EUA	1	165
Itália	1	73
Multinacional	2	715
Portugal	1	217

Fonte: Elaboração própria.

⁵ O índice H é uma métrica proposta por Jorge E. Hirsch (2005) para avaliar a produtividade e o impacto de um pesquisador. Ele é definido pelo número **H**, onde o autor tem **H** artigos com pelo menos **H** citações cada. Esse indicador equilibra quantidade e impacto da produção científica, sendo amplamente utilizado em avaliações e sites de referências acadêmicas (como o Google Acadêmico).

Já o Quadro 2 traz a listagem de revistas nominalmente, indicando: 1) quantas foram as publicações da revista no período delimitado; 2) quantas dessas publicações são recuperadas na busca coluna “Nº Tema na Busca”), e 3) quantos dos textos recuperados possuem os termos buscados entre suas palavras-chaves, seu resumo ou em seu corpo (etapa de checagem texto a texto, assinalada na coluna “Nº Tema na Checagem”). Os dados permitem calcular a proporção de textos sobre IA em relação ao total publicado por cada revista.

Quadro 2 - Volume de textos publicados vs volume de textos no tema da pesquisa

N.	Revista Ativa	Nacionalidade	Nº Textos Total (1)	Nº Tema na Busca (2)	Nº Tema na Checagem (3)	Proporção (3) / (1) %
1	ACB	Brasil	202	0	0	0
2	Acervo	Brasil	204	2	2	0,98
3	Ágora	Brasil	147	3	2	1,36
4	Archeion Online	Brasil	118	1	1	0,85
5	AtoZ	Brasil	160	0	0	0
6	Biblionline	Brasil	141	0	0	0
7	Biblos	Brasil	140	0	0	0
8	BRAJIS	Brasil	182	4	1	0,55
9	IBICT	Brasil	158	0	0	0
10	Em Questão	Brasil	334	2	0	0
11	Encontros Bibli	Brasil	204	2	2	0,98
12	InCID	Brasil	123	0	0	0
13	Informação Arquivística	Brasil	19	0	0	0
14	Informação & Informação	Brasil	301	1	1	0,33
15	Informação & Sociedade	Brasil	119	2	1	0,84
16	Liinc em Revista	Brasil	211	1	1	0,47
17	Perspectivas em Ciência da Informação	Brasil	178	0	0	0
18	PG&C	Brasil	216	0	0	0
19	PontodeAcesso	Brasil	108	0	0	0
20	RDBCI	Brasil	149	0	0	0
21	Revista do AGCRJ	Brasil	88	0	0	0
22	Revista do APEES	Brasil	94	1	0	0
23	Revista do Arquivo	Brasil	134	0	0	0
24	Revista Ibero-Americana de Ciência da Informação	Brasil	224	3	3	1,34

25	Revista Sociais e Humanas - CCSH	Brasil	71	0	0	0
26	Transinformação	Brasil	109	0	0	0
27	<i>American Archivists</i> ⁶	EUA	165	8 ⁷	4	2,42
28	<i>Archival Science</i>	Multinacional	93	12	11	11,83
29	<i>Archivamos</i>	Espanha	208	9	6	2,88
30	<i>Archivaria</i>	Canadá	83	2	1	1,2
31	<i>El Profesional de la Información</i>	Multinacional	622	0	0	0
32	<i>Nuovi Annali della Scuola Speciale per Archivisti e Bibliotecari</i>	Itália	73	0	0	0
33	<i>Páginas A&B</i>	Portugal	217	1	1	0,46
34	<i>Tábula</i>	Espanha	65	2	2	3,08
TOTAL			5660	56	39	0,69

Fonte: Elaboração própria.

Das 34 revistas ativas na Etapa 2, apenas 17 recuperaram dados na busca às palavras-chaves, somando 56 textos para a análise individualizada da Etapa 3, apenas 0,95% dos textos totais (5660). Na Etapa 3, o número de textos é reduzido mais uma vez, pela falta de acesso livre a alguns e pela constatação da ausência dos termos “IA” e “arquivologia” no resumo, entre as palavras-chaves e no corpo do texto ou menção apenas tangencial (como em descrição de campo de tabela, ou em seção “sobre o autor”)⁸. Ao final, resultam 15 revistas e 39 textos para a análise da Etapa 4 (Quadro 3).

Quadro 3 - Distribuição dos textos na Etapa 4

Ref. Arquivologia	Tema tangencial (IA)	Tema central (IA)
Sim	22	12
Não	4	1
Total	25	13

Fonte: Elaboração própria.

Na Etapa 5, para responder à problemática (se há exemplos práticos de uso da IA nas publicações recentes de Arquivologia), considerou-se prática de aplicação a descrição da IA como uma ferramenta de funcionalidade específica, ou quando sua abordagem dentro de um campo específico de atuação na Ciência da Informação; os

⁶ Em cinza, as revistas estrangeiras.

⁷ Presença de textos sem acesso livre (3 em Archivamos, 1 em American Archivist).

⁸ As menções são tangenciais em 5 textos: 1 na revista *Ágora*, 1 na *BRAJIS*, 1 na *APEES* e 2 em *The American Archivist*.

demais casos foram rotulados “menção geral”. Dessa análise qualitativa, foram filtradas 18 entradas que, além desses critérios, fazem menção à Arquivologia.

Quadro 4 - Textos de IA como prática e informações gerais de autoria

N.	Periódico	Nacionalidade Revista	Título	Ano	Nº Cit	Autor	Índice H Autor	País Autor
1	Ágora	Brasil	<i>Relato de experiência: InterPARES Summer school</i>	2024	0	Luciane B. Adolfo	NA	Brasil
2	<i>Archival Science</i>	Multinacional	<i>How can we make born-digital and digitised archives more accessible?</i>	2022	24	Lise Jaillant	9	RU
3	<i>Archival Science</i>	Multinacional	<i>Archivist in the machine: paradata for AI-based automation in the archives</i>	2023	3	Jeremy Davet	NA	Canadá
						Babak Hamidzadeh	NA	EUA
						Patricia Franks	10	EUA
4	<i>Archival Science</i>	Multinacional	<i>The perpetual twilight of records</i>	2024	0	Gregory Rolan	10	Austrália
						Antonina Lewis	5	Austrália
5	<i>Archival Science</i>	Multinacional	<i>The cloud, the public square, and digital public archival infrastructure</i>	2023	0	Tom Nesmith	NA	Canadá
6	<i>Archival Science</i>	Multinacional	<i>Archives, linked data and the digital humanities</i>	2022	32	Ashleigh Hawkins	NA	RU
7	<i>Archival Science</i>	Multinacional	<i>Finding values, building communities</i>	2024	0	Naya Sucha-xaya	1	Tailândia
8	<i>Archival Science</i>	Multinacional	<i>Record DNA</i>	2023	3	Julie McLeod	14	RU
						Elizabeth Lomas	NA	RU
9	<i>Archivamos</i>	Espanha	<i>ChatGPT en modo archivero</i>	2023	2	Diego R.Gutiérrez	NA	Espanha
10	<i>Archivamos</i>	Espanha	<i>Datos sintéticos para inteligencias artificiales</i>	2022	0	Pilar Campos	NA	Espanha
11	<i>Archivamos</i>	Espanha	<i>¿Sueñan las inteligencias artificiales con archivos personales?</i>	2022	0	Pilar Campos	NA	Espanha
12	<i>Archivamos</i>	Espanha	<i>Inteligencia artificial y archivos, de reto a realidad</i>	2021	0	Lluís-Esteve Casellas	3	Espanha
13	<i>Archivamos</i>	Espanha	<i>Inteligencia artificial en archivos</i>	2021	0	Diego Rodríguez Gutiérrez	NA	Espanha

14	Tábula	Espanha	<i>Inteligencia artificial en los archivos. Consideraciones de diseño e implementación</i>	2022	0	Alicia B. Amozurrutia	NA	México
						Yaminel B. Astorga	2	México
						Rodrigo C. Hidalgo	NA	México
						Claudia A. E. Velázquez	NA	México
						Luis M. García-Velázquez	3	México
15	Tábula	Espanha	<i>IA y archivo - La tecnología de reconocimiento de textos manuscritos en fondos patrimoniales</i>	2022	1	Pepita R. Pajares	NA	Espanha
						Celio H. Tornero	3	Espanha
						Meritxell S. Martin	5	Espanha
16	<i>The American Archivist</i>	EUA	<i>More Data, Less Process</i>	2022	2	Lise Jaillant	9	RU
17	<i>The American Archivist</i>	EUA	<i>Maintaining Records in Context?</i>	2020	11	Ciaran B. Trace	11	EUA
18	<i>The American Archivist</i>	EUA	<i>Responsible Operations: Data Science, Machine Learning, and AI in Libraries</i>	2021	71	Thomas Padilla	NA	EUA
						Rachel MacGregor	NA	RU

Fonte: Elaboração própria.

Além de apresentar os títulos dos textos colhidos na análise qualitativa, o Quadro 4 traz informações relativas a suas autorias⁹, indicando o ano de lançamento do texto, seu volume de citações (coluna “Nº Cit”) e Índice H dos autores (ambos retirados do Google Acadêmico) e os países de origem deles. A partir dos dados do Quadro 4, outras inferências se fazem possíveis acerca dos 18 textos finais.

Quanto ao Índice H, fica claro, no quadro, que o volume de citações não é diretamente impactado pela relevância do autor (medida pelo índice). É o que se vê no caso dos autores Patricia Franks e Gregory Rolan, ambos com Índice H no valor 10 e textos sem citação. Contudo, o fato de autores bastante referenciados (caso de Julie McLeod e Ciaran B. Trace, com Índice H superior a 10) possuírem pesquisas sobre o tema demonstra um interesse crescente em torno dele.

⁹ Devido à limitação no número de páginas para a apresentação completa das referências ao final do trabalho, optou-se por incluir no Quadro 4 o máximo de informações referentes aos textos recuperados, garantindo, assim, a transparência e a rastreabilidade dos dados analisados.

Dos 18 textos, 9 não possuem citação, implicando 50% da amostragem. É uma taxa alta, mas deve ser considerada com ressalvas: primeiro, porque o tempo recortado nesta pesquisa é de quatro anos (2020-2024), de modo que o material selecionado é produção recente. Além disso, os 9 restantes perfazem 149 citações sobre o tema, um volume expressivo no período de um quadriênio. Há textos que, neste período, foram citados mais de 10 vezes, sendo destaques os de Thomas Padilla e Rachel MacGregor (2021), com 71 citações; de Ashleigh Hawkins (2022), com 32; de Lise Jaillant (2022), com 24, e de Ciaran B. Trace (2020), com 11. Embora muitos dos textos elencados não versem exclusivamente sobre IA, há que se descartar 1) a importância que ela assume nas propostas acerca do documento digital e 2) sua presença crescente nas discussões sobre o ambiente virtual e seu futuro.

Colocando em escala as publicações encontradas, nota-se um aumento do tema (IA como prática) nas produções. Mesmo o volume de 2024 - cuja amostra colhida encerrou em março/2024 - é igual ao total de 2021 e superior ao de 2020.

Gráfico 1- Publicações por Ano



Fonte: Elaboração própria.

Os 18 textos analisados são de 5 revistas e 26 autores, de diferentes países. Nos Gráficos 2 e 3, há um retrato dessas distribuições.

Gráfico 2 - Distribuição de Revistas por País

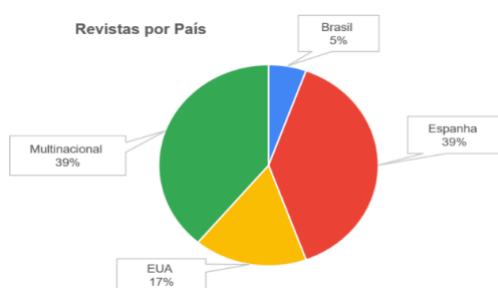
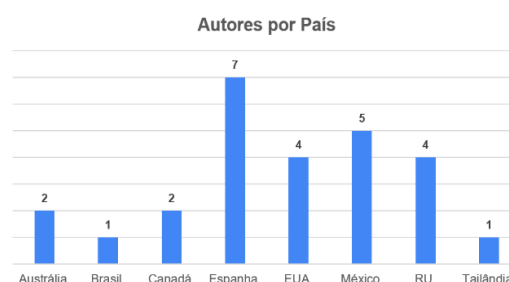


Gráfico 3 - Distribuição de Autores por País



Fonte: Elaboração própria (2025)

Ressalta-se que, apesar do montante inicial das revistas ser composto, em sua maior parte, de títulos brasileiros, elas quase não figuram na lista final. Das 34 revistas totais elencadas na Etapa 2 (Quadro 2), o país responde por 26 (76,5%), com 4134 publicações entre os 5660 totais (73%). Desse montante de publicações, apenas 1 revista (Ágora) e 1 texto compôs a amostragem final (Etapa 5), no Quadro 4. Portanto, é 1 entrada frente 4134: 0,024% do total das publicações nacionais. Em comparação, os dados das multinacionais têm um desempenho superior: são 8 revistas em 34 (23,5%), com 715 publicações em 5660 (12,6%). No Quadro 4, contudo, respondem por 17 dos 18 textos indicados. Um título multinacional sozinho, *Archival Science*, é responsável por 7 das 18 publicações (38,9%). São 7 para 715 entradas: 0,98%. No cenário geral, a Espanha sagrou-se campeã: no período avaliado, as duas publicações mencionadas respondem por um total de 273 textos. Desses, como se vê no Quadro 4, 7 falam de IA e mencionam Arquivologia - 2,5% de seu volume editorial.

4 CONSIDERAÇÕES FINAIS

O campo amostral da pesquisa concentrou-se em revistas conhecidas no campo da Arquivologia, indicadas em *sites* de língua portuguesa. Essa delimitação é relevante, pois, ao mesmo tempo em que assegura a análise de publicações de referência no Brasil, restringe o escopo e exclui, dissertações, teses, livros e outras revistas não contempladas, incluindo publicações em outros idiomas. Ainda assim, trata-se de um recorte válido, pois abrange fontes de ampla circulação e impacto na produção acadêmica nacional.

Neste sentido, os resultados obtidos são significativos, ao refletir o que chega ao público arquivístico brasileiro no que se refere à presença e à prática da inteligência artificial. Do total de 56 textos identificados no período analisado (2020-2024), 18 tratam a IA como uma ferramenta prática e fazem menção direta à Arquivologia. Mesmo em uma leitura inicial, observa-se que todos esses textos abordam a tecnologia relacionando as práticas e os desafios específicos da gestão de documentos, apontando formas de automação em processos como organização, indexação e recuperação de documentos arquivísticos, entre outros. Em paralelo, destaca-se a recorrência de reflexões sobre as implicações éticas da IA, especialmente no que se refere à preservação, ao consumo da

informação e ao papel do arquivista e demais profissionais da informação como mediadores dessas práticas.

As abordagens relevam uma pluralidade de perspectivas sobre como a integração da IA à Arquivologia, desde sua aplicação direta até as implicações mais amplas para a profissão. Os dados indicam um interesse crescente pelo tema da IA e seu uso, mas que ainda não se traduz em expressividade nas publicações do campo, sobretudo no contexto brasileiro. A presença do tema, no período, está praticamente restrita aos textos estrangeiros, nos quais vem aumentando.

A diferença observada entre a produção nacional e estrangeira aponta para um descompasso que não pode ser ignorado. A baixa incidência de textos brasileiros que tratam da aplicação prática da IA na Arquivologia pode refletir tanto a tímida adoção dessas tecnologias no cotidiano profissional quanto um atraso na incorporação crítica do tema no debate acadêmico. Em ambos os casos, evidencia-se uma lacuna que precisa ser superada. Avançar na produção científica, estimular pesquisas empíricas e promover a integração responsável das tecnologias emergentes são passos fundamentais para o fortalecimento da Arquivologia e sua atuação diante dos desafios impostos pela transformação digital.

REFERÊNCIAS

KURZWEIL, R. **Como criar uma mente**: os segredos do pensamento humano. São Paulo: Editora Aleph, 2015.

LÉVY, P. **Cibercultura**. São Paulo: Editora 34, 1999.

LOPES, G. Revistas científicas Arquivologia e C.I. **Arquivista.blog.br**. Disponível em: <https://arquivista.blog.br/http-arquivista-blog-br-revistas-cientificas-com-foco-na-arquivologia/>. Acesso em: 08 abr. 2025.

PERIÓDICOS científicos. **Associação dos arquivistas do estado do Rio de Janeiro (AAERJ)**. Disponível em: <https://aaerj.org.br/links/periodicos-cientificos/>. Acesso em: 08 abr. 2025.

TAYLOR, H. A. Transformation in the archives: technological adjustment or paradigm shift? **Archivaria**, Vancouver, v. 25, p. 12–28, 1988. Disponível em: <https://archivaria.ca/index.php/archivaria/article/view/11451>. Acesso em: 08 abr. 2025.

Universidade Federal de Santa Maria (UFSM). **Arquivologia**: Periódicos da área. Disponível em: <https://www.ufsm.br/cursos/graduacao/santa-maria/arquivologia/publicacoes>. Acesso em: 08 abr. 2025.

WEISS, M. C. Sociedade sensoriada: a sociedade da transformação digital. **Cidade e ambiente Estud.**, São Paulo, v. 33, n. 95, p. 203-214. 2019. Disponível em: <http://dx.doi.org/10.1590/s0103-4014.2019.3395.0013>. Acesso em: 08 abr. 2025.



[4.0 Internacional](https://creativecommons.org/licenses/by/4.0/)

Esta licença permite compartilhamento, remixe, adaptação e criação a partir do trabalho, mesmo para fins comerciais, desde que sejam atribuídos créditos ao(s) autor(es). Conteúdos elaborados por terceiros, citados e referenciados nesta obra não são cobertos pela licença.

Como citar este trabalho:

GUERREIRO, Gabriela de Grande; LATGÉ, Ana Maria dos Santos; GAVA, Tânia Barbosa Salles; ALEIXO, Diana Vilas Boas Souto. Arquivologia e inteligência artificial: análise das publicações e tendências no período 2020-2024. *In*: WORKSHOP DE INFORMAÇÃO, DADOS E TECNOLOGIA, 8., Marília, SP. **Anais [...]**. Marília, SP: Universidade de Marília. 2025. DOI: <https://doi.org/10.22477/viii.widat.242>.



WIDaT 2025

VIII WORKSHOP DE INFORMAÇÃO,
DADOS E TECNOLOGIA

MARÍLIA • SP

Integração da Inteligência Artificial no mapeamento arquivístico: perspectivas iniciais

Integration of Artificial Intelligence in archival mapping: initial perspectives

Maria Fabiana Izídio de Almeida¹



<http://lattes.cnpq.br/7246492153479687>



<https://orcid.org/0000-0001-5602-2274>

Fernando Alves da Gama²



<http://lattes.cnpq.br/1630147171680159>

Fernanda Furio Crivellaro³



<http://lattes.cnpq.br/7099638321605432>



<https://orcid.org/0000-0002-5024-2900>

Eliandro dos Santos Costa⁴



<http://lattes.cnpq.br/7189992612241818>



<https://orcid.org/0000-0001-9475-1574>

Resumo

A Arquivologia tem o desafio de acompanhar a evolução tecnológica, a fim de propor soluções que abarque desde a preservação da informação, até mesmo garantir sua autenticidade nos mais variados suportes e aparatos tecnológicos. A velocidade e volume de dados, documentos e informações colaboram para esse cenário desafiador, no qual a gestão de documentos está posicionada e precisa ser implementada de modo eficaz. Para que isso ocorra, é necessário um trabalho preliminar para conhecer o funcionamento da organização. Destaca-se o mapeamento arquivístico como metodologia apropriada para o levantamento dessas informações. Sabe-se que a Inteligência Artificial (IA) tem se mostrado eficiente em produzir novos conteúdos, refinar dados, tendo em vista os repertórios e catálogos já inseridos na sua base de dados. Tendo em vista o cenário exposto coloca-se como questão-problema desta pesquisa: Em quais etapas do

¹ Doutora em Ciência da Informação, Universidade Estadual Júlio Mesquita Filho (UNESP), Marília, São Paulo, Brasil. Professora Colaboradora do Departamento de Ciência da Informação, Universidade Estadual de Londrina (UEL), Londrina, Paraná, Brasil. maria.fabiana@uel.br.

² Mestre em Ciência da Informação, Universidade Estadual Júlio Mesquita Filho (UNESP), Marília, São Paulo, Brasil. Arquivista em órgão público no Estado do Paraná. Curitiba, Paraná, Brasil. feralgama@gmail.com.

³ Doutora em Ciência da Informação, Universidade Estadual Júlio Mesquita Filho (UNESP), Marília, São Paulo, Brasil. Coordenadora do curso de Administração, Fundação Universitária para o Desenvolvimento do Ensino e da Pesquisa (FUNDEPE), Penápolis, São Paulo, Brasil. fernanda.crivellaro1@gmail.com.

⁴ Doutor em Ciência da Informação, Universidade Estadual Júlio Mesquita Filho (UNESP), Marília, São Paulo, Brasil. Professor do Departamento de Ciência da Informação, Universidade Estadual de Londrina (UEL), Londrina, Paraná, Brasil. eliandro.costa@uel.br.

mapeamento arquivístico a Inteligência Artificial pode contribuir? Para responder essa questão traçou-se como objetivo geral desta pesquisa: Identificar em quais atividades do Mapeamento arquivístico a Inteligência Artificial pode contribuir, a fim de otimizá-las. Esta pesquisa é de natureza qualitativa, do tipo exploratória e o estudo desenvolveu-se a partir de pesquisa bibliográfica. Aponta-se como principal resultado a contribuição da IA com cinco atividades do Mapeamento arquivístico: Pesquisar a história da organização, história administrativa, história dos fundadores e história do arquivo; identificar e pesquisar o ramo de negócio; identificar a missão da organização; mapear processos e elaborar relatório das atividades do Mapeamento arquivístico.

Palavras-chave: arquivologia; mapeamento arquivístico; Inteligência Artificial; Inteligência Artificial Generativa.

Abstract

Archival Science faces the challenge of keeping pace with technological advancements in order to propose solutions that encompass everything from information preservation to ensuring its authenticity across various media and technological devices. The speed and volume of data, documents, and information contribute to this challenging scenario, in which records management is positioned and needs to be implemented effectively. To achieve this, a preliminary effort is required to understand the functioning of the organization. Archival mapping stands out as an appropriate methodology for gathering this information. It is known that Artificial Intelligence (AI) has proven effective in generating new content and refining data, considering the repertoires and catalogs already integrated into its database. Given the outlined scenario, the central research question is: At which stages of archival mapping can Artificial Intelligence contribute? To answer this question, the general objective of this research was outlined as: To identify the activities within Archival Mapping where Artificial Intelligence can contribute in order to optimize them. This research is of a qualitative nature, exploratory in type, and the study was developed through bibliographic research. The main result points to AI's contribution to five activities of Archival Mapping: Researching the organization's history, administrative history, founders' history, and archival history; Identifying and researching the business sector; Identifying the organization's mission; Mapping processes; and Preparing a report on the archival mapping activities.

Keywords: archival science; archival mapping; Artificial Intelligence; Generative Artificial Intelligence.

1 INTRODUÇÃO

A Ciência da Informação ao longo da sua história recebeu em diferentes momentos reflexos da globalização e da revolução tecnológica, isso porque abarca em seu objeto de estudo a informação, logo “[...] é um campo dedicado às questões científicas e à prática profissional voltadas para os problemas da efetiva comunicação do conhecimento e de seus registros entre os seres humanos [...]” (Saracevic, 1996, p. 47). Para a Arquivologia, não é diferente, além de mudanças profundas nos próprios paradigmas, tem o desafio de acompanhar a evolução tecnológica, a fim de propor soluções que abarquem desde a preservação da informação, até mesmo garantir autenticidade das informações produzidas nos mais variados suportes e aparatos tecnológicos.

A velocidade e volume de dados e informações colaboram para esse cenário desafiador, no qual a gestão de documentos está posicionada, pois é trata-se de uma ferramenta indispensável para a otimização do uso das informações contidas nos mais variados suportes, ou seja, é o trabalho de assegurar que a informação orgânica seja gerenciada com economia e eficácia, para que possa ser facilmente recuperada, de modo ágil e fornecendo subsídios para as ações das organizações com respostas esclarecidas, rápidas, seguras, confiáveis e que permitam a redução do nível de incertezas, visando a obtenção da excelência na realização das atividades diárias das organizações (Moreno, 2008).

Para que isso ocorra, é necessário um trabalho preliminar, árduo e fundamental para conhecer o funcionamento da organização, para além das condições físicas dos ambientes informacionais. Destaca-se o mapeamento arquivístico como metodologia apropriada para o levantamento dessas informações, pois consiste “[...] na pesquisa, na identificação, no levantamento e no controle de todas as variáveis que refletem no arquivo da organização [...]” (Almeida, 2023, p. 173), logo contempla elementos informacionais, que abarcam o planejamento e ações da gestão de documentos.

Dentre as etapas propostas no mapeamento arquivístico é preciso mapear os processos, os quais foram identificadas lacunas, dificuldades e ou oportunidades. Como resultado desse mapeamento de processos são elaborados os fluxogramas, tal qual o processo se realiza e a partir disso, propõe-se um novo fluxo, com todas as melhorias visualizadas. Nesse momento, é crucial utilizar tecnologias, que compilem, sistematizem os dados coletados e forneçam também soluções.

Sabe-se que a Inteligência Artificial (IA) tem se mostrado eficiente em produzir/novos conteúdos, refinar dados, tendo em vista os repertórios e catálogos já inseridos na sua base de dados. Tendo em vista o cenário exposto coloca-se como questão-problema desta pesquisa: Em quais etapas do mapeamento arquivístico a Inteligência Artificial pode contribuir? Para responder essa questão traçou-se como objetivo geral desta pesquisa: identificar em quais atividades do mapeamento arquivístico a Inteligência Artificial pode contribuir, a fim de otimizá-las.

Esta pesquisa é de natureza qualitativa, do tipo exploratória e o estudo desenvolveu-se a partir de pesquisa bibliográfica. Aponta-se como principal resultado a contribuição da IA com cinco atividades do Mapeamento arquivístico: Pesquisar a história da organização, história administrativa, história dos fundadores e história do arquivo; identificar e pesquisar o ramo de negócio; identificar a missão da organização; mapear processos e elaborar relatório das atividades do Mapeamento arquivístico.

1.1 Mapeamento Arquivístico

O contexto de uma organização pública ou privada pode ser compreendido por meio das suas atividades-meio e atividades-fim, pois estas necessitam registrar suas atividades e procedimentos administrativos e jurídicos para manter a documentação disponível para seu uso e em alguns casos para a sociedade. Todas as informações orgânicas produzidas relacionam-se à missão e objetivo de existência da respectiva organização.

Para manter essas informações organizadas e implementar uma gestão de documentos capaz de contribuir com o negócio da empresa e impulsionar a gestão da informação é preciso realizar o mapeamento arquivístico, a fim de identificar as variáveis que refletem no arquivo, isso quer dizer que é necessário identificar questões físicas, documentais e informacionais de toda a organização.

O Mapeamento arquivístico surge *a priori* da necessidade de ampliar o termo diagnóstico de arquivo (Camargo; Bellotto, 2010), uma vez que tem como foco as condições físicas do arquivo, que são altamente relevantes para a organização e prospecção dos acervos, no entanto, no quesito de coletar informações para embasar a gestão de documentos e informações tornou-se inviável para os desafios atuais, pois “[...] a carência de literatura em torno do diagnóstico de arquivo e a própria complexidade dessa atividade, tem levado os profissionais a resultados parciais, incompletos e ineficientes no ato da intervenção” (Almeida; Vitoriano, 2018, p. 81).

Nesse sentido, o Mapeamento arquivístico é uma proposta conceitual e metodológica que fornece ferramentas adequadas para identificar todas as variáveis que refletem no arquivo, tendo em vista questões documentais, informacionais e físicas. O

Mapeamento arquivístico é composto por 15 atividades, que realizado de modo organizado e sistemático proporciona um riquíssimo reportório que embasará o plano de classificação, contribuirá com a elaboração da tabela de temporalidade, auxiliará na elaboração do registro topográfico e ainda contribuirá com a melhoria dos processos da organização. O Quadro 1 – “Atividades do Mapeamento arquivístico” condensa essas atividades e objetivos.

Quadro 1 – Atividades do Mapeamento arquivístico

Atividade	Objetivo
1. Pesquisar a história da organização, história administrativa, história dos fundadores e história do arquivo	Conhecer e identificar a evolução da organização, tendo em vista a organização, administração, fundadores e arquivo
2. Identificar e pesquisar o ramo de negócio	Conhecer o ramo de negócio, uma vez que reflete diretamente nas informações produzidas e recebidas
3. Identificar a missão da organização	Conhecer a missão para identificar as atividades-fim ligadas ao negócio
4. Analisar a estrutura administrativa	Levantar informações que demonstrem como a empresa está organizada estruturalmente
5. Mapeamento de atividades e documentos	Levantar as informações referentes aos processos, atividades e documentos produzidos e recebidos no âmbito de cada processo
6. Levantar as funções e atividades	Levantar informações que demonstrem como a empresa está organizada funcionalmente
7. Analisar o mapeamento de atividades e documentos	Analisar as informações coletadas na atividade 5
8. Elaborar a Lista Mestra do Mapeamento	Relacionar as funções, processos, atividades e documentos produzidos e recebidos
9. Entrevistar os colaboradores	Investigar as lacunas existentes nos processos
10. Identificar os fluxos informacionais	Investigar os fluxos informacionais
11. Mapear processos	Mapear os processos que possuem lacunas, dificuldades e melhorias
12. Realizar o diagnóstico de arquivo	Levantar informações referente às condições físicas dos acervos

13. Realizar o mapeamento documental	Identificar os documentos acondicionados nos acervos
14. Acrescentar informações na Lista Mestra de Mapeamento	Acrescentar os tipos de documentos identificados no acervo na Lista Mestra
15. Elaborar relatório das atividades do Mapeamento arquivístico	Apresentar os resultados do mapeamento arquivístico à direção do projeto

Fonte: Elaboração própria, baseada em Almeida (2023).

Nota-se que o mapeamento arquivístico é uma metodologia que propicia uma densa coleta de dados, que pode variar na proporção do porte da organização. Nesse sentido, é fundamental apoiar-se em recursos tecnológicos capazes de contribuir com esse processo.

1.2 Inteligência Artificial: Contextualização

A Inteligência Artificial (IA) é um dos campos mais recentes e dinâmicos da engenharia e da ciência da computação desenvolvido ao longo das últimas décadas, marcando avanços, desafios e revoluções tecnológicas significativas nesse período.

Para Lopes, Santos e Pinheiro (2014) a IA é constituída de vários procedimentos computacionais, cujas funções realizadas seriam consideradas inteligentes se fossem executadas por seres humanos. Russel e Norvig (2013), por sua vez, apontam que um dos principais objetivos da IA consiste em ensinar os computadores a fazer o que os humanos fazem de melhor, e aprender é sem dúvida a mais importante dessas tarefas. Os autores também dividem a compreensão da IA em quatro campos complementares de abordagem: Pensar como seres humanos, automatizando atividades cognitivas; Pensar racionalmente, com o emprego de modelos racionais em processos de percepção, raciocínio e ação; Agir como seres humanos, executando funções que exigem inteligência humana e que anteriormente só poderiam ser realizadas por pessoas; Agir racionalmente por meio de agentes de software inteligentes em sistemas computacionais (Russel; Norvig, 2013).

Russel e Norvig (2013) observam que os fundamentos científicos da IA são profundamente interdisciplinares, integrando conhecimentos de diversas áreas

científicas como a Filosofia, a Matemática, a Economia, a Neurociência, a Psicologia, a Engenharia da Computação, a Cibernética, Linguística, etc; integrando elementos teóricos ali presentes em prol da criação de agentes de *software* inteligentes. A IA, portanto, não se limita à criação de máquinas inteligentes, mas também busca compreender e replicar a inteligência em suas múltiplas dimensões, abrindo caminho para inovações que transformam a sociedade e a maneira como interagimos com a tecnologia. Para isso, é salutar apresentar, ainda que de modo sucinto, excertos da evolução da IA.

Para compreender a evolução da Inteligência Artificial (IA), é importante reconhecer que essa disciplina se consolidou como uma área de estudo e aplicação prática após a Segunda Guerra Mundial. Desde então, a IA passou por diferentes períodos que foram fundamentais para sua consolidação e expansão. A seguir, apresentamos um breve cronograma destacando os principais marcos históricos que moldaram o desenvolvimento dessa área.

Russel e Norvig (2013) indicam que o primeiro trabalho reconhecido no campo da IA foi realizado em 1943 por Warren McCulloch e Walter Pitts, que propuseram um modelo de neurônios artificiais. Em 1950, Alan Turing introduziu outro marco significativo para a IA, conhecido como o Teste de Turing. Esse teste estabelece que um computador pode ser considerado inteligente se um interrogador humano não conseguir distinguir se as respostas recebidas vêm de uma máquina ou de outro ser humano.

O ano de 1956 é considerado o marco inicial da IA como campo de estudo, quando o termo "Inteligência Artificial" foi inicialmente estabelecido. Em 1958, McCarthy desenvolveu a linguagem Lisp, que se tornou a principal linguagem de programação para IA. Em 1962, Frank Rosenblatt apresentou o teorema da convergência do perceptron, fundamental para o desenvolvimento de redes neurais (Russel; Norvig, 2013).

Desenvolvendo os conceitos metodológicos até ali delineados, o programa DENDRAL representou no ano de 1969 o primeiro sistema bem-sucedido de conhecimento intensivo, marcando a aplicação prática da IA em problemas complexos (Russel; Norvig, 2013).

A década de 1970 é tida por muitos como o "inverno da IA", marcando um período de dificuldades em que os sistemas não atenderam às expectativas e o

entusiasmo que os pesquisadores propagavam nos anos iniciais. Não obstante, a década de 1980 trouxe um renascimento impulsionado por avanços teóricos e computacionais, com o ressurgimento das pesquisas no âmbito das redes neurais. Desse modo, já no ano de 1987, a IA foi reconhecida como uma ciência independente, com aplicações práticas em diversos setores da indústria.

Nos anos 1990, o conceito de agentes inteligentes emergiu, definindo entidades que percebem e agem em seus ambientes de forma autônoma. Essa nova perspectiva focou na interação entre sistemas inteligentes e seus ambientes, enquanto avanços em aprendizado de máquina e reconhecimento de padrões começaram a ganhar destaque (Russel; Norvig, 2013).

A partir dos anos 2000, o avanço da internet e a disponibilidade de grandes volumes de dados (*Big Data*) impulsionaram a IA. A computação em nuvem e o aumento do poder de processamento permitiram o desenvolvimento de técnicas sofisticadas, como redes neurais profundas (*Deep Learning*). Na década de 2010, o *Deep Learning* dominou o campo da IA, com aplicações em processamento de linguagem natural, como os modelos Word2Vec e *Transformers*, tais como GPT, BERT. O aprendizado por reforço também ganhou destaque e popularidade, exemplificado pelo *AlphaGo* da *DeepMind*. Isso permitiu avanços significativos em tradução automática, geração de texto; possibilitando a disponibilidade de assistentes virtuais, como Siri e Alexa (Russel; Norvig, 2013; Lopes; Santos; Pinheiro, 2014; Silva; Peres; Boscarioli, 2016).

Observa-se, na década de 2020, uma explosão massiva no desenvolvimento e no uso de aplicações computacionais da IA altamente promissoras. São exemplos as áreas da IA Multimodal, focada na integração de diferentes tipos de dados (texto, imagem, áudio etc.) para melhorar a capacidade de compreensão e interação dos sistemas de IA; e da IA Quântica, que emprega a computação quântica para potencializar a capacidade de processamento e resolver problemas complexos que exigem cálculos em grande escala (Alayrac et al., 2022; Baltrušaitis; Ahuja; Morency, 2017; Lloyd; Weedbrook, 2018).

Isso potencializa o surgimento de soluções robustas no campo da IA Generativa, que se concentra na criação de novos conteúdos, como textos, imagens, músicas ou vídeos, a partir de dados existentes. A evolução da Inteligência Artificial, desde seus

primórdios até a era contemporânea, reflete um contínuo avanço nas técnicas e abordagens utilizadas para a construção de soluções computacionais inteligentes na resolução de problemas complexos, minimização de tarefas repetitivas, dentre outras funcionalidades que visam otimizar diversos ramos de negócios.

A generalização das Redes **Neurais Artificiais (RNAs)** é um conceito que se refere à habilidade da rede para produzir saídas adequadas em entradas de dados que não estejam presentes durante o treinamento (Lopes; Santos; Pinheiro, 2014; Haykin, 2008). No **Aprendizado de Máquina**, segundo Lopes, Santos e Pinheiro (2014), é a linha de pesquisa em *Machine Learning*, busca desenvolver métodos computacionais que possibilitem a criação de sistemas capazes de aprender de maneira autônoma, priorizando-se a eficiência e o crescimento da capacidade das soluções frente ao conceito de generalização. Na **Mineração de dados** é um processo analítico, automático ou semiautomático, que visa explorar grandes bases de dados para descobrir padrões relevantes que possam embasar a assimilação de informações significativas e apoiar a produção de conhecimento (Silva; Peres; Boscardioli, 2016).

No que se refere o **Processamento de Linguagem Natural (PLN)** desempenha um papel crucial nas análises documentais de IA, garantindo que os sistemas inteligentes extraiam valor em grandes volumes de dados textuais, sendo um dos componentes indispensáveis para a área da Representação do Conhecimento (Mueller; Massaron, 2019). A **Representação do Conhecimento** é considerada por Brachman e Levesque (2004) uma das áreas centrais da IA e a definem como o processo de captura e estruturação do conhecimento, de modo que um sistema computacional possa utilizá-lo na resolução de problemas. Já a **Mineração de Processos** é uma disciplina do campo da IA que aplica as metodologias aqui apresentadas no campo de atuação do Mapeamento, Análise e Modelagem de Processos de Negócios. Aalst (2011) observa que os sistemas computacionais da atualidade registram massivas quantidades de dados sobre eventos, contemplando informações detalhadas sobre as atividades executadas nas organizações.

Nesse cenário, a Inteligência Artificial Generativa (IA Generativa) promove um aprofundamento na integração de redes neurais, *machine learning*, mineração de dados e processamento de linguagem natural (NLP), formando um corpo metodológico capaz

de gerar dados novos e realistas. Nesse contexto, observa-se que evolução das redes neurais artificiais (RNAs) levou ao desenvolvimento de técnicas avançadas, como as Redes Generativas Adversariais (GANs), que geram dados de alta qualidade; combinando modelos probabilísticos, aprendizado não supervisionado, *transformers* e técnicas de difusão, a IA Generativa se consolida como uma ferramenta versátil para análise, geração e otimização em diversas áreas (Goodfellow *et al.* 2014). Para Aruda (2024, p. 4) “A capacidade de tratamento de grandes quantidades de dados é outro aspecto importante a ser considerado no uso de IA generativa [...]”. O tratamento inicial de dados realizado pela IA Generativa, certamente, otimiza diferentes trabalhos, reduzindo custos operacionais e até mesmo recursos humanos.

Com tantas áreas e possibilidades de interação com IA, é latente a necessidade do campo da Ciência da Informação, tendo em vista a Arquivologia, que haja um aprofundamento nas tecnologias atuais e emergentes da IA buscando melhorias que agreguem na performance das suas metodologias e técnicas, visando viabilizar a eficiência dos seus resultados.

2 PROCEDIMENTOS METODOLÓGICOS

Esta pesquisa se caracteriza de natureza qualitativa, do tipo exploratória, uma vez que há lacunas de conhecimento produzido e sistematizado sobre os assuntos aqui propostos. A pesquisa foi desenvolvida no âmbito de pesquisa bibliográfica, nas Bases de Dados de Teses e Dissertações (BDTD), na Base de Dados em Ciência da Informação (BRAPCI) e na *SciELO*. A pesquisa ocorreu no mês de fevereiro de 2025, não foram estabelecidos padrão temporal e não foram utilizados termos booleanos. Os termos de pesquisa adotados foram “inteligência artificial” e “mapeamento arquivístico”. Os resultados obtidos auxiliaram na construção do *corpus* teórico desta pesquisa. Além dos periódicos supracitados, outras obras também foram relevantes para a pesquisa.

3 RESULTADOS

A Inteligência Artificial e sua capacidade de aprender e gerar novos resultados é, sem dúvida, uma revolução posta a toda a sociedade e aproveitar esse recurso é algo almejado por toda área do conhecimento e muitos experimentos têm ocorrido nesse sentido. A partir da literatura levantada em torno da IA nesta pesquisa, chegou-se ao entendimento, que a IA generativa contribui com as atividades do Mapeamento arquivístico, otimizando algumas atividades, pois para entregar um resultado eficiente necessita em cada atividade levantar uma gama de informações, que em dado momento é necessário compilá-las, analisá-las, a fim de propor soluções e sustentar os instrumentos arquivísticos inerentes à gestão de documentos e gestão da informação. Considerando as atividades de pesquisa, análise de informações, elaboração de fluxogramas (Atividade Mapear processos) e elaboração de relatórios, apontam-se a seguir as etapas do Mapeamento arquivístico que a IA Generativa pode contribuir, conforme apresentado no Quadro 2 – “Mapeamento arquivístico e a IA Generativa”, que segue abaixo.

Quadro 2 – Mapeamento arquivístico e a IA Generativa

Atividade	IA Generativa
1. Pesquisar a história da organização, história administrativa, história dos fundadores e história do arquivo	Pode contribuir com pesquisa preliminar, considerando as informações disponíveis na <i>Internet</i>
2. Identificar e pesquisar o ramo de negócio	Após identificar o ramo de negócio, a IA pode contribuir com pesquisa preliminar, considerando as informações disponíveis na <i>Internet</i>
3. Identificar a missão da organização	Pode contribuir com pesquisa preliminar, considerando as informações disponíveis na <i>Internet</i>
11. Mapear processos	Após o mapeamento do processo a IA pode elaborar o diagrama do processo
15. Elaborar relatório das atividades do Mapeamento arquivístico	A IA pode elaborar um texto preliminar ou refinar o relatório do Mapeamento arquivístico, a partir de <i>prompts</i>

Fonte: Elaboração própria.

Nota-se que foi possível identificar cinco atividades do Mapeamento arquivístico que a IA pode contribuir, o que de fato otimiza o tempo da aplicação da metodologia em questão. Dentre as atividades, destaca-se a atividade mapear processos, haja vista que desenhar que uma atividade densa, que a depender da complexidade do processo, pode onerar quaisquer projetos, por outro lado, é uma atividade fundamental que elucida lacunas, dificuldades e oportunidades.

Por meio do mapeamento de processos, é possível estudar e validar cada atividade que faz parte de um processo complexo, os envolvidos passam a ter uma visão ampliada de seus respectivos papéis funcionais, permitindo conhecimento sobre o todo, (Pradella; Furtado; Kipper, 2012). O mapeamento de processo, quando feito adequadamente, proporciona um trabalho contínuo e padronizado, garante excelência e eficiência. Logo, é essencial elaborar os fluxogramas, que são diagramas que representam a sequência de etapas de um processo utilizando símbolos gráficos (Oliveira, 2007). Para que a IA realize os diagramas é preciso ter plataformas que apresentem esse recurso, saber conduzi-la através de *prompts* assertivos e contextualizados para o um resultado esperado.

4 CONSIDERAÇÕES FINAIS

Buscar ferramentas e áreas de conhecimento que colaboram com os desafios postos à Ciência da Informação é emergente. Tendo em vista a Arquivologia e a gestão de documentos encontrar na IA Generativa recursos que impulsionam as atividades e as tornem mais ágeis é muito contributivo. A integração da Inteligência Artificial (IA) no mapeamento arquivístico representa um avanço no campo da gestão e organização de informações. Através do uso de tecnologias avançadas, como a IA generativa é possível transformar a maneira como as informações são coletadas e processadas, permitindo maior automatização e eficiência no processo, acessibilidade e pesquisa de informações.

Esta pesquisa, pode-se dizer de caráter preliminar, identificou que a IA contribui diretamente em cinco atividades do Mapeamento arquivístico: na pesquisa da história da organização, história administrativa, história dos fundadores e história do arquivo; na identificação e pesquisa o ramo de negócio; na identificação a missão da organização; no mapeamento de processos e na elaboração de relatório das atividades do Mapeamento arquivístico. Pretende-se estender esse estudo para uma pesquisa de campo, a fim de identificar outras possibilidades de uso e aplicação da IA.

REFERÊNCIAS

- AALST, Wil M. P. van der. **Process mining**: discovery conformance and enhancement of business processes. Berlin: Springer, 2011.
- ALMEIDA, Maria F. I. de; VITORIANO, Marcia C. de C. P. Diagnóstico de arquivos e mapeamento da informação: interlocução da gestão documental com a gestão da informação. **Em Questão**, [S. l.], v. 24, n. 3, 2018. Disponível em: <https://brapci.inf.br/v/10040>. Acesso em: 20 jan. 2025.
- ALAYRAC, Jean-Baptiste *et al.* Flamingo: a visual language model for few-shot learning. **ArXiv**, [S. l.], 2022. Disponível em: <https://arxiv.org/abs/2204.14198>. Acesso em: 23 fev. 2025.
- ALMEIDA, Maria Fabiana Izídio de. **Mapeamento Arquivístico**: uma proposta conceitual e metodológica para ambientes organizacionais. Tese (Doutorado) – Faculdade de Filosofia e Ciências (FFC), Universidade Estadual Paulista (Unesp), Marília, 2023. Disponível em: <https://repositorio.unesp.br/entities/publication/7977e2e4-6b55-4be0-9dd6-52d5302061f7>. Acesso em: 28 fev. 2025.
- ARUDA, Eucidio Pimenta. Inteligência Artificial Generativa no contexto da transformação do trabalho docente. **Educação Em Revista**, Belo Horizonte, v. 40, e48078, 2024. Disponível em: <https://doi.org/10.1590/0102-469848078>. Acesso em: 23 fev. 2025.
- BALTRUŠAITIS, Tadas; AHUJA, Chaitanya; MORENCY, Louis-Philippe. Multimodal Machine Learning: a survey and taxonomy. **ArXiv**, [S. l.], p. 1-40, 2017. Disponível em: <https://arxiv.org/pdf/1705.09406>. Acesso em: 15 fev. 2025.
- BRACHMAN, Ronald; LEVESQUE, Hector. **Knowledge representation and reasoning**. San Francisco: Morgan Kaufmann, 2004.
- CAMARGO, Ana Maria de Almeida; BELLOTTO, Heloísa Liberalli. **Dicionário de terminologia Arquivística**. São Paulo: Centro de Memória da Educação FEUSP/FAPESP, 2010.
- GOODFELLOW, Ian J. *et al.* Generative adversarial nets. **Advances in neural information processing systems**, [Montreal], v. 27, 2014. Disponível em: <https://doi.org/10.48550/arXiv.1406.2661>. Acesso em: 25 fev. 2025.
- HAYKIN, Simon. **Redes neurais**: princípios e prática. Porto Alegre: Bookman, 2008.
- LLOYD, Seth; WEEDBROOK, Christian. Quantum generative adversarial learning. **Physical review letters**, [S. l.], v. 121, n. 4, e040502, 2018. Disponível em: <https://arxiv.org/abs/1804.09139>. Acesso em: 10 fev. 2023.
- LOPES, Isaías Lima; SANTOS, Flávia Oliveira; PINHEIRO, Carlos. **Inteligência Artificial**. Rio de Janeiro: GEN LTC, 2014.

MORENO, Nádina Aparecida. Gestão documental ou gestão de documentos: trajetória histórica. In: BARTALO, Linete; MORENO, Nádina Aparecida (org.). **Gestão em Arquivologia**: abordagens múltiplas. Londrina: Editora Edue, 2008.

MUELLER, John Paul.; MASSARON, Luca. **Aprendizado de máquina para leigos**. Rio de Janeiro: Alta Books, 2019.

OLIVEIRA, Djalma de Pinho Rebouças. **Sistemas, organização & métodos**: uma abordagem gerencial. São Paulo: Atlas, 2007.

PRADELLA, Simone; FURTADO, João Carlos; KIPPER, Liane Mahlmann. **Gestão de processos da teoria a prática**. São Paulo: Atlas, 2012.

RUSSEL, Stuart; NORVIG, Peter. **Inteligência Artificial**. 3. ed. Rio de Janeiro: GEN LTC, 2013.

SARACEVIC, Tefko. Ciência da Informação: origem, evolução e relações. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 1, n. 1, p. 41-62, 1996. Disponível em: <https://periodicos.ufmg.br/index.php/pci/article/view/22308>. Acesso em: 9 mar. 2025.

SILVA, Leandro Augusto; PERES, Sarajane Marques.; BOSCARIOLI, Clódis. **Introdução à mineração de dados com aplicações em R**. Rio de Janeiro: GEN LTC, 2016.



[4.0 Internacional](https://creativecommons.org/licenses/by/4.0/)

Esta licença permite compartilhamento, remixe, adaptação e criação a partir do trabalho, mesmo para fins comerciais, desde que sejam atribuídos créditos ao(s) autor(es). Conteúdos elaborados por terceiros, citados e referenciados nesta obra não são cobertos pela licença.

Como citar este trabalho:

ALMEIDA, Maria F. I. de et al. Integração da Inteligência Artificial no mapeamento arquivístico: perspectivas iniciais. In: WORKSHOP DE INFORMAÇÃO DADOS E TECNOLOGIA, 8., 2025, Marília, SP. **Anais [...]**. Marília, SP: Universidade de Marília, 2025. DOI: <https://doi.org/10.22477/viii.widat.261>.



WIDaT 2025

VIII WORKSHOP DE INFORMAÇÃO,
DADOS E TECNOLOGIA

MARÍLIA • SP

Possíveis contribuições de uma ferramenta de Inteligência Artificial para a gestão de documentos

Possible contributions of an Artificial Intelligence tool to document management

Jayron Viana dos Santos¹



<http://lattes.cnpq.br/4934852611370647>



<https://orcid.org/0009-0006-2536-3387>

Natália Marinho do Nascimento²



<http://lattes.cnpq.br/8152394509168605>



<https://orcid.org/0000-0002-0770-2035>

Resumo

Introdução: A Arquivologia, aliada à Ciência da Informação, constitui-se em campo passível de utilização das ferramentas da Inteligência Artificial, sob variados aspectos e processos arquivísticos. Objetiva-se então, compreender como uma ferramenta de IA, especificamente a plataforma *huggingface.co*, pode contribuir para a gestão de documentos, com foco nas funções arquivísticas. Metodologia: Este trabalho, caracteriza-se como estudo de caso, de natureza qualitativa do tipo descritiva e exploratória. Utilizou-se como instrumento de coleta e análise a ferramenta de Inteligência Artificial *huggingface.co*, instruída mediante *prompts* (comandos textuais), verificando como esta ferramenta compreende as funções arquivísticas da gestão de documentos, indicando ainda outras ferramentas de Inteligência Artificial com possíveis contribuições. Resultados: As análises e resultados acerca do uso da ferramenta *Huggingface.co*, mostraram-se bastante satisfatórios e com contribuições relevantes, mas ainda assim, alertando-se para o fato de se utilizar essas ferramentas de IA com a devida orientação e acompanhamento, incluindo sobretudo os aspectos éticos. Conclusão: Considera-se que apesar da ferramenta *Huggingface.co* recuperar conteúdo relevante no que tange a gestão de documentos na Arquivologia ainda carece de maior embasamento científico e prático em seus usos, necessitando do acompanhamento de profissionais das mais variadas especialidades atuando em parceria para a interpretação e treinamento da ferramenta.

Palavras-chave: inteligência artificial; gestão de documentos; funções arquivísticas.

Abstract

Introduction: Archival Science, combined with Information Science, constitutes a field capable of using Artificial Intelligence tools, under various archival aspects and processes. The objective, therefore, is to understand how an AI tool, specifically the platform *huggingface.co*, can contribute to document management, with a focus on archival functions. Methodology: This work is characterized as bibliographical research, of a qualitative, descriptive and exploratory nature. The Artificial Intelligence tool *huggingface.co* was used as a collection and analysis instrument, instructed through prompts (textual commands), verifying how this tool understands the archival functions of record management, also indicating other Artificial Intelligence tools with possible

¹ Mestrando, Programa de Pós-Graduação em Ciência da Informação (PPGCI) da Unesp, Marília, São Paulo, Brasil. Professor Efetivo do Curso de Bacharelado em Biblioteconomia, Universidade Estadual do Piauí (UESPI), Teresina, Piauí, Brasil. jayron.viana@unesp.br.

² Doutora em Ciência da Informação, Programa de Pós-Graduação em Ciência da Informação (PPGCI) da Unesp, Marília, São Paulo, Brasil. Professora Assistente, Faculdade de Filosofia e Ciências (FFC), Marília, São Paulo, Brasil. natalia.nascimento@unesp.br.

contributions. Results: The results regarding the use of the Huggingface.co tool proved to be quite satisfactory and with relevant contributions, but still, alerting to the fact of using these AI tools with due guidance and monitoring, including above all ethical aspects. Conclusion: It is considered that although the Huggingface.co tool retrieves relevant content regarding document management in Archival Science, it still requires greater scientific and practical foundation in its applications, necessitating the guidance of professionals from various specialties to collaborate in the interpretation and training of the tool.

Keywords: artificial intelligence; record management; archival functions.

1 INTRODUÇÃO

Estudos recentes no campo da Ciência da Informação (CI) têm explorado os avanços da inteligência artificial (IA) e suas potenciais contribuições em diversos contextos, destacando-se o interesse mútuo entre ambas as áreas no que diz respeito ao tratamento da informação, visando facilitar o cotidiano dos usuários. A CI desempenha um papel fundamental na organização, classificação e recuperação de informações, garantindo a qualidade e relevância dos dados que alimentam os sistemas de IA, essenciais para a construção de novos conhecimentos.

Além disso, a CI contribui para a definição de políticas e diretrizes éticas no uso da IA, abordando questões como integridade, privacidade e proteção de dados, bem como a mitigação de vieses algorítmicos, assegurando a transparência e confiabilidade das informações geradas (Moreira *et al.*, 2023).

O desenvolvimento acelerado de tecnologias de IA tem impactado significativamente diversas áreas, incluindo a Arquivologia, ao otimizar processos de gestão documental. Conforme Santos (2021), a gestão de documentos envolve um conjunto de procedimentos que visam a racionalidade e eficácia no tratamento de informações, desde sua produção até o arquivamento, promovendo o acesso ágil e organizado.

Nesse contexto, esta pesquisa busca compreender como uma ferramenta de IA, especificamente a plataforma *huggingface.co*, pode contribuir para a gestão de documentos, com foco nas funções arquivísticas. O estudo analisa como essa ferramenta interpreta e aplica comandos específicos (*prompts*) para otimizar processos como produção, tramitação, avaliação e conservação de documentos, além de identificar possíveis intervenções que aprimorem essas atividades.

A pesquisa caracteriza-se como qualitativa, descritiva e exploratória, utilizando um estudo de caso para examinar as funcionalidades da *huggingface.co* e suas respostas a comandos específicos. A aproximação entre IA e Arquivologia é relevante diante dos avanços

tecnológicos e da necessidade de agilidade e assertividade nas atividades e etapas de gestão documental.

Assim, o estudo visa demonstrar como ferramentas de IA podem ser integradas às práticas arquivísticas, promovendo maior eficiência e inovação no tratamento da informação.

1.1 Perspectivas da IA e a Arquivologia

Os avanços da Inteligência Artificial (IA), impulsionados pelo progresso tecnológico, resultaram na criação de diversas tecnologias com impacto direto no cotidiano (Coneglian *et al.*, 2023). Entre suas principais subáreas destacam-se o processamento de linguagem natural, aprendizado de máquina e profundo, sistemas especialistas, robótica e Big Data, que seguem em evolução (Gontijo; Araújo, 2021).

A IA tem origem nos estudos de Alan Turing, na década de 1950, que associou a computação à ideia de inteligência futura. Em 1950, ele propôs o "Teste de Turing", questionando a capacidade das máquinas de pensar. Em 1966, Joseph Weizenbaum desenvolveu ELIZA, um dos primeiros chatbots, demonstrando o potencial da IA para interação com humanos (Gunkel, 2017).

No contexto digital, o crescente volume de dados nato-digitais e digitalizados exige novas estratégias para sua identificação, acesso e organização. O tratamento digital demanda padrões e protocolos específicos, e a IA desempenha um papel central, oferecendo mecanismos de interoperabilidade e automação na estruturação da informação (Martins, 2010).

As principais funções da IA incluem aprendizagem, raciocínio, planejamento e processamento de linguagem, sendo operacionalizadas pelo aprendizado de máquina (ML) e aprendizado profundo (DL) (Russel; Norvig, 2013). O ML treina algoritmos para tomada de decisões com base em dados, enquanto o DL utiliza redes neurais para analisar grandes volumes de informações, como imagens e texto (Janiesh; Zschechi; Heinrich, 2021).

O processamento de linguagem natural (PLN) é essencial para a extração e organização de dados textuais, podendo adotar abordagens simbólicas, estatísticas, conexionistas ou híbridas (Bulegon; Moro, 2010). Na Arquivologia, a IA contribui para a transparência dos registros, o acesso à informação e a recuperação de dados. Técnicas como ML auxiliam na

classificação de documentos, enquanto o PLN facilita a extração de informações relevantes (Felipe, 2023; Dorneles, 2022).

Diante disso, é essencial que arquivistas compreendam a IA e desenvolvam competências na área para garantir a conformidade das tecnologias aplicadas à gestão documental. Apesar do potencial da IA, sua aplicação ainda está concentrada na fase inicial da gestão documental, tornando sua implementação um desafio contínuo (Rolan *et al.*, 2019). Os avanços tecnológicos desse modo, exigem que os profissionais da área acompanhem essas mudanças para assegurar a eficiência da gestão arquivística no futuro.

1.2 Gestão de Documentos e Funções Arquivísticas

A gestão de documentos é fundamental para assegurar o tratamento, armazenamento, organização e disseminação adequados dos documentos de arquivo, independentemente de seu suporte. Após a Segunda Guerra Mundial, o aumento exponencial na produção de informações demandou novas soluções para gerenciar o volume crescente de documentos, consolidando a gestão de documentos como um conceito essencial, alinhado à arquivística e à teoria das três idades (Flores; Lampert, 2013).

Segundo Nascimento (2019), a gestão de documentos atende a necessidades administrativas e organizacionais, refletindo a própria organização ao registrar suas atividades e trajetória. Sua existência influencia processos organizacionais e a sociedade, promovendo transparência e boa governança. Entre seus objetivos, destacam-se a agilização do acesso à informação, a promoção da transparência, a eficiência administrativa, o controle do fluxo documental e a racionalização da produção de documentos (Bernardes; Delatorre, 2008).

Machado e Camargo (1999) definem a gestão de documentos como um conjunto de procedimentos técnicos que abrangem a produção, tramitação, uso, avaliação e arquivamento de documentos. A negligência nessa gestão pode resultar em fragmentação de sistemas de informação, perda de contextualidade e declínio na qualidade dos objetos de informação, comprometendo sua procedência, integridade e preservação (Van Bussel, 2017).

A ausência de um programa de gestão documental pode levar à perda de autenticidade dos documentos, especialmente ao atingirem seu valor secundário. A interoperabilidade entre sistemas é crucial para rastrear a trajetória dos documentos,

impactando diretamente a preservação digital. Nesse contexto, as funções de produção, classificação e avaliação são essenciais. A produção está ligada às demandas administrativas e ao valor primário dos documentos (administrativo, legal ou fiscal). A classificação organiza os documentos em classes, enquanto a avaliação determina prazos de guarda e eliminação, resultando em planos de classificação e tabelas de temporalidade (Heredia Herrera, 1993).

Portanto, a gestão de documentos, ao garantir o respeito aos valores primário e secundário, bem como a autenticidade e integridade dos documentos, é fundamental para a manutenção da cadeia de custódia e impacta todas as funções arquivísticas, incluindo aquelas relacionadas aos arquivos permanentes.

2 PROCEDIMENTOS METODOLÓGICOS

Este trabalho adotou uma abordagem qualitativa, que, conforme Creswell (2007), é essencialmente interpretativa, permitindo ao pesquisador analisar os dados a partir de narrativas complexas e interativas, que proporcionam uma compreensão mais abrangente do objeto de estudo.

Além disso, foram utilizadas as metodologias de pesquisa exploratória e descritiva. A pesquisa exploratória possibilita uma análise sistêmica dos dados qualitativos, facilitando uma interpretação detalhada do fenômeno investigado (Losh; Rambo; Ferreira, 2023). Já a pesquisa descritiva consiste na observação, registro, análise e organização dos dados, sem intervenção direta do pesquisador (Prodanov; Freitas, 2013).

O estudo de caso foi aplicado como método de investigação, seguindo a perspectiva de Yin (2015), que destaca sua capacidade de ir além da descrição de fatos, buscando proporcionar conhecimento sobre o fenômeno estudado e validar ou contrastar relações identificadas, e tendo como objetivos explorar, descrever, explicar, avaliar e transformar, sendo o foco deste trabalho a ferramenta de IA *Huggingface.co*.

Para embasar a pesquisa, realizou-se um levantamento bibliográfico em bases de dados, livros, teses e dissertações, abordando temas como IA, CI, Arquivologia, Gestão de Documentos e funções arquivísticas. Em seguida, utilizou-se a técnica de observação para coletar dados sobre o funcionamento da ferramenta *Huggingface.co*, com o intuito de compreender sua aplicação no processo de gestão de documentos, especialmente em relação às funções arquivísticas.

Os testes com a ferramenta foram conduzidos no final de junho e início de julho de 2024, registrados por meio de *prints* de tela e posteriormente analisados. A análise dos dados seguiu o método indutivo, que, conforme Prodanov e Freitas (2013), parte da observação de fatos ou fenômenos, buscando identificar relações entre eles e, a partir dessas conexões, proceder à generalização dos resultados observados. Essa abordagem permitiu uma compreensão aprofundada das possíveis contribuições da ferramenta de IA para a gestão de documentos e funções arquivísticas.

3 RESULTADOS

A *Hugging Face* é uma empresa especializada em Inteligência Artificial (IA), destacando-se pelo desenvolvimento de ferramentas de *Machine Learning* e Processamento de Linguagem Natural (PLN). Sua principal contribuição é a biblioteca de código aberto *Transformers*, que disponibiliza modelos pré-treinados de PLN e recursos para treinamento e implementação de modelos personalizados, além de *datasets* acessíveis à comunidade (Pareto, 2024).

A escolha do chatbot *Huggingface.co* baseia-se em sua capacidade de indicar ferramentas de IA e suas aplicações em áreas especializadas, como a Arquivologia. Embora menos popular que o *ChatGPT*, que lidera globalmente com 14,6 bilhões de visitas anuais, o *Huggingface.co* ocupa a quinta posição, com cerca de 316,6 milhões de acessos no mesmo período.

Para avaliar seu funcionamento, foram aplicados comandos em linguagem natural, analisando sua compreensão, estrutura, pertinência e assertividade das respostas, bem como sua aplicabilidade na gestão documental mediada por IA e rigor científico das fontes recuperadas.

Quadro 1 - *Huggingface.co*: comandos e respostas obtidas ³

COMANDO 1	COMANDO 2	COMANDO 3
Você agora é um arquivista! Cite e caracterize cada uma das funções	Você agora é um consultor de gestão de documentos e utiliza	Você é um arquivista! Relacione as funções arquivísticas e

² Os *datasets* (ou conjuntos de dados) são o principal insumo dos processos de análise de dados. Disponível em: <https://aquare.la/datasets-o-que-sao-e-como-utiliza-los/>.

³ Disponível em: <https://huggingface.co/docs/transformers/pt/index>.

arquivísticas e de que maneira cada uma dessas funções pode ser auxiliada por ferramentas de inteligência artificial	ferramentas de Inteligência Artificial na gestão arquivística! Quais ferramentas de Inteligência Artificial você utilizaria em seu trabalho?	ferramentas de Inteligência Artificial que auxiliam em cada uma dessas funções.													
RESPOSTA 1	RESPOSTA 2	RESPOSTA 3													
De acordo com Rousseau e Couture (1998), as funções arquivísticas são divididas em sete categorias, cada uma descrevendo uma atividade essencial para a gestão eficiente de arquivos e documentos. Aqui está uma lista dessas funções e de que maneira a Inteligência Artificial (IA) pode auxiliar em cada uma delas: 1.Produção/Criação: Essa função diz respeito à geração de novos documentos e registros. A IA pode auxiliar na produção de documentos por meio da automatização da criação de conteúdo, geração de modelos e até mesmo na criação de novos formatos, como vídeos e áudios interativos, que podem enriquecer o acervo. 2.Avaliação: A avaliação envolve determinar o valor dos documentos, sua autenticidade e relevância para a coleção. Ferramentas de IA podem agilizar esse processo ao analisar grandes volumes de dados simultaneamente, identificando padrões e atribuindo valores com base em critérios pré-estabelecidos. 3. Aquisição: A aquisição trata da coleta e recebimento de documentos para o acervo. A IA pode auxiliar na identificação de fontes relevantes de informação, monitorando publicações e atividades online, bem como sugerindo aquisições com base em critérios definidos. 4.Conservação/Preservação: A conservação visa garantir a integridade física e digital dos documentos ao longo do tempo. A IA pode ajudar a preservar os arquivos digitais, sugerindo métodos de armazenamento seguro e eficiente, monitorando a integridade dos dados e até mesmo restaurando arquivos danificados.	Como consultor em gestão documental, eu utilizaria uma variedade de ferramentas de Inteligência Artificial (IA) para otimizar os processos de gestão arquivística e melhorar a organização e acessibilidade dos documentos. Algumas ferramentas de IA que eu empregaria incluem: 1.Reconhecimento Óptico de Caracteres (OCR): Essa tecnologia permite que o texto de documentos digitalizados ou escritos manualmente seja reconhecido e convertido em dados digitais pesquisáveis. Isso agiliza a busca e recuperação de informações específicas dentro de documentos. 2.Processamento de Linguagem Natural (PLN): Essa tecnologia pode extrair informações relevantes, resumir textos, identificar sentimentos e relações entre conceitos, facilitando a classificação e indexação de documentos. 3.Aprendizado de Máquina (Machine Learning, ML): Essa tecnologia pode ser usada para detectar padrões em grandes conjuntos de documentos, prever categorias de classificação e até mesmo identificar documentos anômalos ou confidenciais que requerem atenção especial. 4.Automação Robótica de Processos (RPA): A RPA envolve a automação de tarefas repetitivas, como o roteamento de documentos, a extração de dados de formulários e a migração de arquivos entre sistemas. 5.Análise Preditiva: Ferramentas de IA podem usar análise preditiva para prever a relevância e o valor de um documento ao longo do tempo. Isso pode ajudar na priorização de documentos para preservação ou	Claro! Aqui está um quadro que relaciona as funções arquivísticas com possíveis ferramentas de Inteligência Artificial que podem auxiliar em cada uma delas: <table><tr><th>Função Arq.</th><th>Ferramenta</th></tr><tr><td>1. Produção</td><td>IA de Geração de Conteúdo -ChatGPT -Midjourney</td></tr><tr><td rowspan="2">2. Avaliação</td><td>IA de PLN -Google Cloud Natural Language -Azure Cognitive Services, da Microsoft -IBM Watson</td></tr><tr><td>IA de Visão Computacional -TensorFlow -PyTorch -OpenCV</td></tr><tr><td rowspan="2">3. Aquisição</td><td>Web Scraping IA -Selenium -Beautiful Soup -API especializadas (Oxylabs Scraper API)</td></tr><tr><td>IA de reconhecimento de padrões - Scikit-learn - TensorFlow - PyTorch</td></tr><tr><td rowspan="2">4.Conservação</td><td>IA de Preservação Digital -Preservica - Arkivum</td></tr><tr><td>IA de Monitoramento Ambiental</td></tr></table>	Função Arq.	Ferramenta	1. Produção	IA de Geração de Conteúdo -ChatGPT -Midjourney	2. Avaliação	IA de PLN -Google Cloud Natural Language -Azure Cognitive Services, da Microsoft -IBM Watson	IA de Visão Computacional -TensorFlow -PyTorch -OpenCV	3. Aquisição	Web Scraping IA -Selenium -Beautiful Soup -API especializadas (Oxylabs Scraper API)	IA de reconhecimento de padrões - Scikit-learn - TensorFlow - PyTorch	4.Conservação	IA de Preservação Digital -Preservica - Arkivum	IA de Monitoramento Ambiental
Função Arq.	Ferramenta														
1. Produção	IA de Geração de Conteúdo -ChatGPT -Midjourney														
2. Avaliação	IA de PLN -Google Cloud Natural Language -Azure Cognitive Services, da Microsoft -IBM Watson														
	IA de Visão Computacional -TensorFlow -PyTorch -OpenCV														
3. Aquisição	Web Scraping IA -Selenium -Beautiful Soup -API especializadas (Oxylabs Scraper API)														
	IA de reconhecimento de padrões - Scikit-learn - TensorFlow - PyTorch														
4.Conservação	IA de Preservação Digital -Preservica - Arkivum														
	IA de Monitoramento Ambiental														

5.Classificação: A classificação envolve organizar e categorizar os documentos de acordo com um sistema lógico. A IA pode agilizar esse processo ao analisar o conteúdo dos documentos e atribuir tags e categorias automaticamente, facilitando a recuperação posterior. 6.Descrição: A descrição cria registros que representam os documentos de forma acurada e acessível. A IA pode extrair informações-chave dos documentos e gerar descrições e metadados automaticamente, economizando tempo e garantindo consistência. 7.Difusão: A difusão visa disponibilizar os documentos para acesso e uso pelo público. A IA pode auxiliar na difusão ao tornar os documentos mais acessíveis por meio de interfaces de pesquisa intuitivas, tradução de idiomas e até mesmo sugerindo conteúdos relevantes com base nas necessidades dos usuários. A IA oferece oportunidades significativas para melhorar a eficiência e a eficácia das práticas arquivísticas. Ao automatizar tarefas repetitivas e fornecer insights valiosos, os arquivistas podem se concentrar em atividades mais complexas e estratégicas, melhorando a gestão e acessibilidade dos acervos. FONTES: proad.ufba.br;agora.emnuvens.com.br;arquivistica.fci.unb.br;exposicao-virtual-gestao-de-servicosarquivisticos.webnode.page .	descarte, garantindo que os acervos permaneçam relevantes e gerenciáveis. 6.Sistemas de Recomendação: Esses sistemas de IA podem sugerir documentos relevantes com base nas preferências e comportamentos dos usuários, melhorando a experiência do usuário e aumentando a probabilidade de descoberta de informações valiosas. 7.Segurança Baseada em IA: Ferramentas de IA podem ser usadas para proteger os documentos confidenciais contra acesso não autorizado. Isso inclui a detecção de padrões de comportamento suspeito, criptografia baseada em IA e monitoramento contínuo da integridade dos dados. 8.Indexação Automática: A IA pode atribuir tags e metadados automaticamente aos documentos com base em seu conteúdo, tornando a recuperação de informações muito mais rápida e precisa. 9.Monitoramento de Conformidade: Soluções de IA podem ajudar a garantir a conformidade regulatória ao monitorar o conteúdo dos documentos e garantir que eles atendam aos padrões legais, éticos e organizacionais relevantes. Essas ferramentas de IA podem ser integradas a uma plataforma de gestão documental abrangente, que oferece aos arquivistas e usuários finais uma experiência eficiente e intuitiva. A IA revolucionou a gestão documental, melhorando a precisão, a velocidade e a acessibilidade dos arquivos. FONTES: blog.zapsign.com.br;destaque.com.br;unite.ai;updf.com;pt.linkedin.com	<table><tr><td></td><td>Sensera Systems-ThingsLog</td></tr><tr><td rowspan="2">5.Classificação</td><td>IA de Classificação de Conteúdo - Microsoft Share Point - Google Cloud Vision API - IBM Watson</td></tr><tr><td>IA de Indexação Automática - Elasticsearch - Apache Solr</td></tr><tr><td rowspan="2">6. Descrição</td><td>Geração de Metadados IA - Adobe Sensei - IBM Watson</td></tr><tr><td>Extração de Entidades Nomeadas IA -SpaCy -Stanford NER -Google Cloud Natural Language</td></tr><tr><td rowspan="2">7. Difusão</td><td>Recomendações de IA -Amazon Personalize -Microsoft Recommender -TensorFlow Recommender</td></tr><tr><td>IA de pesquisa e Recuperação de informações -Elasticsearch -Solr -Azure Cognitive Search</td></tr></table>		Sensera Systems-ThingsLog	5.Classificação	IA de Classificação de Conteúdo - Microsoft Share Point - Google Cloud Vision API - IBM Watson	IA de Indexação Automática - Elasticsearch - Apache Solr	6. Descrição	Geração de Metadados IA - Adobe Sensei - IBM Watson	Extração de Entidades Nomeadas IA -SpaCy -Stanford NER -Google Cloud Natural Language	7. Difusão	Recomendações de IA -Amazon Personalize -Microsoft Recommender -TensorFlow Recommender	IA de pesquisa e Recuperação de informações -Elasticsearch -Solr -Azure Cognitive Search
	Sensera Systems-ThingsLog												
5.Classificação	IA de Classificação de Conteúdo - Microsoft Share Point - Google Cloud Vision API - IBM Watson												
	IA de Indexação Automática - Elasticsearch - Apache Solr												
6. Descrição	Geração de Metadados IA - Adobe Sensei - IBM Watson												
	Extração de Entidades Nomeadas IA -SpaCy -Stanford NER -Google Cloud Natural Language												
7. Difusão	Recomendações de IA -Amazon Personalize -Microsoft Recommender -TensorFlow Recommender												
	IA de pesquisa e Recuperação de informações -Elasticsearch -Solr -Azure Cognitive Search												
FONTES: tableau.com													

Fonte: Dados de pesquisa (2024).

Com base no quadro analítico e na literatura nacional e internacional sobre gestão de documentos, este estudo buscou analisar como a ferramenta de

IA *Huggingface.co* compreende e relaciona as funções arquivísticas no processo de gestão documental, além de identificar outras tecnologias de IA que possam otimizar essas funções, além do domínio e recuperação de fontes científicas confiáveis acerca das temáticas requeridas. Para isso, foram realizados três comandos específicos, cujas respostas foram analisadas quanto à coerência, pertinência e relevância.

No **Comando 1**, solicitou-se que a ferramenta atuasse como um arquivista, citando e caracterizando as funções arquivísticas e indicando como ferramentas de IA poderiam auxiliar em cada uma delas. A **Resposta 1** demonstrou coerência e precisão, abordando todas as funções arquivísticas e relacionando-as a tecnologias de IA, recuperando uma única fonte considerada científica, referindo-se a Rousseau e Couture (1998).

No **Comando 2**, a ferramenta foi orientada a atuar como um consultor de gestão de documentos, indicando quais ferramentas de IA utilizaria em seu trabalho. A **Resposta 2** recuperou tecnologias como Reconhecimento Óptico de Caracteres (OCR), Processamento de Linguagem Natural (PLN) e Aprendizado de Máquina (ML), descrevendo seu funcionamento e aplicação no contexto da gestão documental. O texto mostrou-se organizado e pertinente, embora tenha utilizado o termo "tecnologia" como sinônimo de "ferramenta".

No **Comando 3**, a ferramenta foi solicitada a relacionar as funções arquivísticas com as ferramentas de IA que poderiam auxiliar em cada uma delas. A **Resposta 3** apresentou um quadro estruturado, associando funções como avaliação, classificação e descrição a tecnologias específicas, como o IBM Watson, destacando sua aplicação em múltiplas funções simultaneamente. A ferramenta mostrou-se intuitiva, adaptando-se ao contexto dos comandos anteriores e recuperando informações ordenadas e relevantes.

No entanto, a análise das fontes utilizadas pela ferramenta revelou limitações. Apenas uma fonte foi considerada científica e de referência na área de Arquivologia (Rousseau; Couture, 1998), enquanto as demais estavam associadas a grupos empresariais, sites e blogs, sem caráter científico. Isso levanta questões sobre a credibilidade e precisão das respostas, uma vez que a IA pode priorizar conteúdos acessíveis na web, sem acessar literatura especializada ou repositórios institucionais. Essa dependência de fontes não científicas pode resultar em informações imprecisas ou desatualizadas, destacando a necessidade de um uso crítico dessas ferramentas.

Em síntese, embora a *Huggingface.co* tenha demonstrado capacidade de organização e coerência nas respostas, sua eficácia depende diretamente da qualidade das fontes consultadas. Portanto, é essencial que os usuários avaliem criticamente as informações geradas, considerando as limitações e potenciais vieses das fontes utilizadas pela IA.

4 CONSIDERAÇÕES FINAIS

Este estudo objetivou apresentar possíveis contribuições do uso de ferramentas da Inteligência Artificial no processo de Gestão de documentos, com foco nas funções arquivísticas envolvidas nesse processo, analisando a compreensão, capacidade de interação, bem como a pertinência e assertividade das respostas e conteúdos recuperados da ferramenta de IA *Huggingface.co* (chatbot).

Com base nas discussões, verificou-se que a ferramenta de IA *Huggingface.co* demonstrou capacidade de compreensão e análise acerca das instruções e comandos a ela direcionados, com relevantes contribuições acerca das possibilidades de uso e aplicações no desenvolvimento das funções arquivísticas e do processo de gestão de documentos como um todo.

E aliado a isso, outro ponto relevante na discussão, gira em torno da “etapa” de “Descrição”, entendida pela ferramenta como uma “função” da gestão de documentos, somente associada à literatura sob essa ótica em discussões mais contemporâneas. Isso pode estar ligado ao surgimento e desenvolvimento dos documentos e ambientes digitais, visto que com a preocupação da autenticidade, da cadeia de custódia e da preservação digital sistêmica, os documentos devem estar armazenados em sistemas adequados e a descrição de metadados preenchida corretamente no momento da produção do documento.

Por fim, cabe ressaltar que, apesar desta ferramenta recuperar conteúdo tido como relevante, deve-se tomar cuidado com os usos de ferramentas de IA, e não associá-las a “soluções milagrosas”, uma vez que a maioria destas soluções/ferramentas de IA ainda dependem de embasamento científico e prático em seus usos, estando isso intimamente relacionado ao acompanhamento de profissionais das mais variadas especialidades que interpretem os resultados e “treinem” intensivamente estas ferramentas, com vistas a usos mais éticos e eficientes, sobretudo no que diz respeito ao domínio e relações entre as mais variadas fontes científicas.

REFERÊNCIAS

- BERNARDES, I. P.; DELATORRE, H. **Gestão documental aplicada**. São Paulo: Arquivo Público do Estado de São Paulo, 2008.
- BULEGON, Hugo; MORO, Claudia Maria Cabral. Mineração de texto e o processamento de linguagem natural em sumários de alta hospitalar. **Journal of Health Informatics**, São Paulo, v. 2, n. 2, 2010. Disponível em: <https://jhi.sbis.org.br/index.php/jhi-sbis/article/view/5>. Acesso em: 01 jul. 2024.
- CONEGLIAN, Caio Saraiva; TORINO, Emanuelle; SANTARÉM SEGUNDO, José Eduardo; VIDOTTI, Silvana Aparecida Borsetti Gregório. Inteligência artificial generativa e recuperação da informação: tendências e oportunidades de pesquisa. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO (ENANCIB), 23., 2023, Aracaju. **Anais [...]**. Aracaju: ANCIB, 2023. Disponível em: <https://conferencias.ancib.org/index.php/enancib/xxxiiienancib/paper/viewFile/1944/1401>. Acesso em: 03 jul. 2024.
- CRESWELL, J. W. **Projeto de pesquisa: métodos qualitativo, quantitativo e misto**. Porto Alegre: Artmed, 2007.
- DORNELES, Sânderson Lopes. Aplicações de Inteligência Artificial na gestão documental: revisão sistemática na Lista, Emerald e Scopus. **Revista Fontes Documentais**, Aracaju, v. 5, 2022. Disponível em: <https://periodicos.ufba.br/index.php/RFD/article/view/57966/31312>. Acesso em: 01 jul. 2024.
- FELIPE, Esther Quêzia dos Santos. **Uso das plataformas de linguagem natural como ferramentas na indexação de arquivos digitais**. Trabalho de conclusão de Curso (Arquivologia) – Universidade Federal da Paraíba, João Pessoa, 2023. Disponível em: <https://repositorio.ufpb.br/jspui/handle/123456789/28344>. Acesso em: 01 jul. 2024.
- FLORES, D.; LAMPERT, S. R. As funções de produção, classificação e avaliação de documentos arquivísticos no software nuxeo document management. **Informação Arquivística**, Rio de Janeiro, v. 2, n. 1, p. 41-64, jan./jun., 2013.
- GONTIJO, M. C. A.; ARAÚJO, R. F. de. Inteligência artificial: impacto acadêmico e social de sua produção científica. **Ciência da Informação Express**, p. 1-6, 2021. Disponível em: <https://cienciadainformacaoexpress.ufla.br/index.php/revista/article/view/29>. Acesso em: 2 jul. 2024.
- GUNKEL, David J. Comunicação e Inteligência Artificial: novos desafios e oportunidades para a pesquisa em comunicação. **Galaxia**, 2017. Disponível em: <https://www.scielo.br/j/gal/a/K7cGL4Jgp8dXYwzDPZPy6qS/?lang=pt&format=pdf>. Acesso em: 30 jun. 2024.
- HEREDIA HERRERA, A. **Archivística general: teoría y práctica**. Espanha: Diputación Provincial de Sevilla, 1993.

JANIESCH, Christian; ZSCHECHI, Patrick; HEINRICH, Kai. Machine learning and deep learning. **Electronic Markets**, v. 31, p. 685-695, 2021. Disponível em: <http://dx.doi.org/10.1007/s12525-021-00475-2>. Acesso em: 03 jul. 2024.

LÖSCH, S.; RAMBO, C. A.; FERREIRA, J. L. A pesquisa exploratória na abordagem qualitativa em educação. **Revista ibero-americana de estudos em educação**, Araraquara, v. 18, 2023. Disponível em: <https://periodicos.fclar.unesp.br/iberoamericana/article/view/17958>. Acesso em: 01 jul. 2024.

MACHADO, H. C.; CAMARGO, A. M. de. **Como implantar arquivos públicos municipais**. São Paulo: Arquivo do Estado, 1999.

MARTINS, Agnaldo Lopes. Potenciais aplicações da inteligência artificial na ciência da informação. **Informação & Informação**, Londrina, v. 15, n. 1, p. 1-16, 2010. Disponível em: <https://ojs.uel.br/revistas/uel/index.php/informacao/article/view/3882>. Acesso em: 5 jul. 2024.

MOREIRA, Jonathan Rosa; NASCIMENTO, Cláudia Pinheiro; RIBEIRO, Jefferson Bruno Pereira. Ciência da informação e Inteligência Artificial: reflexões sobre o ChatGPT e a informação científica. **Revista Tecnologias em Projeção**, v. 14, n. 2, p. 1-13, 2023. Disponível em: <https://revista.faculdadeprojecao.edu.br/index.php/Projecao4/article/view/2244>. Acesso em: 02 jul. 2024.

NASCIMENTO, Natália Marinho do. **Inter-relação entre os fluxos informacionais, a identificação de tipos documentais e a avaliação de documentos: um modelo processual para a Salutar de Marília: Unesp, 2019. Tese (Doutorado) – Universidade Estadual Paulista (Unesp), Marília, SP, 2019.** Disponível em: <https://repositorio.unesp.br/entities/publication/aeb7c8ab-3781-4156-b0fa-4df8617f932c>. Acesso em: 02 jul. 2024.

PARETO. **Hugging face: o que essa empresa faz?** 2024. Disponível em: <https://blog.pareto.io/hugging-face/>. Acesso em: 01 jul. 2024.

PRODANOV, Cleber Cristiano; FREITAS, Ernani César de. **Metodologia do trabalho científico: métodos e técnicas de pesquisa e do trabalho acadêmico**. 2. ed. Novo Hamburgo: Feevale, 2013.

ROLAN, G.; HUMPHRIES G.; JEFFREY, L.; SAMARAS, E.; ANTISOUPOVA, T.; STUART, K. More human than human?: Artificial Intelligence in the archive. **Archives & Manuscripts**, v. 47, n. 2, p. 179-203, 2019. Disponível em: <https://publications.archivists.org.au/index.php/asa/article/view/10311>. Acesso em: 01 jul. 2024.

RUSSELL, Stuart; NORVIG, Peter. **Inteligência Artificial: uma abordagem moderna**. 3. ed. Rio de Janeiro: Elsevier, 2013.

SANTOS, J. C. Dos. Gestão documental e gestão da informação abordagens, modelos e etapas. **Informação@Profissões**, Londrina, v. 10, n. 1, p. 99-120, 2021. Disponível em:

<https://ojs.uel.br/revistas/uel/index.php/infoprof/article/view/43273>. Acesso em: 28 jun. 2024.

TURING, Alan. **Cérebros, máquinas e consciência**: uma introdução à filosofia da mente. São Carlos: EdUFScar, 1996.

VAN BUSSEL, G. J. **The theoretical framework of the 'archive-as-Is**: an organization oriented view on Archives. Amsterdam: University of Applied Sciences. 2017. Disponível em: https://pure.hva.nl/ws/portalfiles/portal/16954086/The_theoretical_framework_of_the_Archive_as_Is_Archival_theories_I.pdf. Acesso em: 02 jul. 2024.

YIN, Robert K. **Estudo de caso**: planejamento e métodos. 5.ed. Porto Alegre: Bookman, 2015. 290 p.



[4.0 Internacional](https://creativecommons.org/licenses/by/4.0/)

Esta licença permite compartilhamento, remixe, adaptação e criação a partir do trabalho, mesmo para fins comerciais, desde que sejam atribuídos créditos ao(s) autor(es). Conteúdos elaborados por terceiros, citados e referenciados nesta obra não são cobertos pela licença.

Como citar este trabalho:

SANTOS, Jayron Viana dos; NASCIMENTO, Natália Marinho do. Possíveis contribuições de uma ferramenta de Inteligência Artificial para a gestão de documentos. *In*: WORKSHOP DE INFORMAÇÃO, DADOS E TECNOLOGIA, 8., Marília, SP. **Anais [...]**. Marília, SP: Universidade de Marília. 2025. DOI: <https://doi.org/10.22477/viii.widat.247>.



WIDaT 2025

VIII WORKSHOP DE INFORMAÇÃO,
DADOS E TECNOLOGIA

MARÍLIA • SP

Preservação digital sistêmica: o início da construção de um modelo conceitual para o E-Docs/ES

Systemic digital preservation: starting to build a conceptual model for E-Docs/ES

Tânia Barbosa Salles Gava¹



<http://lattes.cnpq.br/0822963734544353>



<https://orcid.org/0000-0002-3920-5623>

Diana Vilas Boas Souto Aleixo²



<http://lattes.cnpq.br/5082397017911269>



<https://orcid.org/0000-0003-4156-354X>

Tatiana Canelhas Pignataro³



<http://lattes.cnpq.br/4042865673221060>



<https://orcid.org/0000-0001-8400-6140>

Resumo

Este artigo apresenta os passos iniciais do Projeto de pesquisa intitulado Preservação Digital Sistêmica dos documentos arquivísticos no âmbito do Governo do Estado do Espírito Santo, o qual visa a construção de um modelo conceitual para a Preservação Digital Sistêmica dos documentos arquivísticos digitais (nato digitais e representantes digitais) do Sistema Corporativo de Gestão de Documentos Arquivísticos Digitais (E-Docs), usado pelos órgãos e entidades do Executivo Estadual do Governo do estado do Espírito Santo. Trata-se de uma pesquisa qualitativa, aplicada e exploratória, fundamentada na relevância prática, no aproveitamento do conhecimento prévio e na colaboração de todos os pesquisadores e envolvidos. Como principais resultados apresentam-se as principais metas do projeto, bem como seus resultados parciais. Conclui-se que o projeto se justifica pela relevância da preservação dos documentos arquivísticos do Poder Executivo Estadual de todo o estado, e por produzir conhecimento científico e tecnológico e qualificação profissional na área de Preservação Digital de todos os envolvidos no projeto.

Palavras-chave: preservação digital sistêmica; sistemas informatizados de gestão arquivística de documentos; E-Docs.

¹ Doutora em Engenharia Elétrica - Automação, Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal do Espírito Santo (PPGEE/UFES), Vitória, ES, Brasil. Docente, Departamento de Arquivologia, Universidade Federal do Espírito Santo (CCJE), Vitória, ES, Brasil. tania.gava@ufes.br.

² Doutora em Ciência da Informação, Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual Paulista (PPGCI/UNESP), Marília, SP, Brasil. Docente, Departamento de Arquivologia, Universidade Federal do Espírito Santo (CCJE), Vitória, ES, Brasil. diana.aleixo@ufes.br.

³ Mestranda em Ciência da Informação, Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual Paulista (PPGCI/UNESP), Marília, SP, Brasil. Consultora, META Documentos Digitais, Brasília, DF, Brasil. tatyacs@gmail.com.

Abstract

This article presents the initial steps of the research project entitled *Systemic Digital Preservation of Archival Documents within the Government of the State of Espírito Santo*. The project aims to develop a conceptual model for the *Systemic Digital Preservation* of digital archival records (born-digital and digital representations) within the Corporate System for Digital Archival Document Management (*E-Docs*) used by agencies and entities of the Executive Branch of the Government of Espírito Santo. This is a qualitative, applied, and exploratory research, based on practical relevance, prior knowledge utilization, and collaboration among researchers and stakeholders. The main results include the project's key objectives and preliminary findings. The study concludes that the project is justified by the importance of preserving the archival records of the State Executive Branch, as well as by its contribution to scientific and technological knowledge production and professional qualification in the field of Digital Preservation.

Keywords: systemic digital preservation; information systems management of archival documents; E-Docs.

1 INTRODUÇÃO

Este artigo tem como objetivo apresentar os passos iniciais de um Projeto de Pesquisa intitulado “Preservação Digital Sistêmica dos documentos arquivísticos no âmbito do Governo do Estado do Espírito Santo”, que é uma parceria da Universidade Federal do Espírito Santo (UFES) com o Arquivo Público do Estado do Espírito Santo (APEES) e Instituto de Tecnologia da Informação e Comunicação do Estado do Espírito Santo (Prodest), que tem como principal objetivo desenvolver um modelo conceitual para a preservação digital dos documentos arquivísticos do E-Docs, que é o Sistema Corporativo de Gestão de Documentos Arquivísticos Digitais, usado pelos órgãos e entidades do Executivo Estadual Governo do Estado do Espírito Santo.

Com a evolução da transformação digital nas instituições desde os anos 1990 e com sua aceleração durante a pandemia, a produção de documentos em formato digital tem se tornado predominante em comparação ao suporte em papel. Dada a complexidade e fragilidades associadas aos documentos digitais, surgiu a necessidade de desenvolver e implementar ações para a preservação desse patrimônio arquivístico de forma sistêmica.

Ou seja, deve-se ver a preservação digital não apenas com foco na obsolescência dos suportes, formatos, *softwares* e *hardwares*, mas sim em uma abordagem sistêmica que contempla todo o ciclo de vida dos documentos arquivísticos digitais (Gava; Flores, 2021b), ou seja, deve-se ter uma Preservação Digital Sistêmica (PDS).

No contexto deste estudo, a PDS é compreendida como um modelo de gestão de preservação digital inspirado no Modelo *Open Archival Information System* (OAIS),

amplamente reconhecido como referência internacional para a preservação de documentos digitais a longo prazo. O modelo abrange as etapas de produção, gestão, preservação, acesso e disseminação de documentos digitais.

Desenvolvido pelo *Consultative Committee for Space Data Systems (CCSDS)* da *National Aeronautics and Space Administration (NASA)*, o modelo OAIS possibilitou a elaboração da norma ISO 14721:2003 (Innarelli, 2015). Posteriormente, a norma foi revisada na versão ISO 14721:2012 “*Space data and information transfer systems: open archival information system – Reference model*” (*International Organization for Standardization*, 2012). Atualmente a versão mais nova do Modelo OAIS é o Magenta Book de dezembro de 2024 (Consultative Committee for Space Data Systems, 2024).

No Brasil, o modelo OAIS foi traduzido e publicado pela Associação Brasileira de Normas Técnicas (ABNT) e publicado sob a forma da norma ABNT NBR 15472:2007, com o título “Sistema Aberto de Arquivamento de Informação SAAI”. Em julho de 2021, a ABNT elaborou a norma ABNT NBR ISO 14721:2021 Sistemas espaciais de transferência de dados e de informação - Sistema Aberto de Arquivamento de Informação (SAAI), adotando integralmente, em conteúdo técnico, estrutura e redação, a versão da ISO 14721:2012. A nova norma revogou e substituiu a ABNT NBR 15472:2007 (Dorneles; Corrêa; Flores, 2024).

Neste sentido a PDS requer uma cadeia de custódia ressignificada para o ambiente digital, ou seja, uma Cadeia de Custódia Digital Arquivística. Segundo Gava e Flores (2020, p. 92) o conceito de Cadeia de Custódia Digital Arquivística (CCDA) “[...] deve trazer a ideia de que a cadeia de custódia digital não pode ser interrompida, e deve ser auditada pela cadeia de preservação ou outro procedimento capaz dessa garantia no ambiente digital.”

A CCDA deve perpassar todos os ambientes referentes às três entidades externas do Modelo OAIS (produtor, administrador e consumidor), que são, respectivamente, o ambiente de gestão de documentos, o ambiente de preservação e o ambiente de acesso e difusão. No Brasil, o ambiente de gestão para a esfera executiva deve ser um Sistema Informatizado de Gestão Arquivística de Documentos (SIGAD), que é orientado pelo Modelo E-Arq Brasil (Conselho Nacional de Arquivos, 2020). O ambiente de preservação, também para o público do executivo, deve ser um Repositório Arquivístico Digital

Confiável (RDC-Arq), cuja implementação é orientada pela Resolução nº 51 do Conselho Nacional de Arquivos (2023). Embora não tenha uma norma consolidada, como as citadas anteriormente, no Brasil, o ambiente de acesso e difusão deve buscar atender normas como a Norma Geral Internacional de Descrição Arquivística (ISAD-G) (Conselho Internacional de Arquivos, 2000), a Norma Brasileira de Descrição Arquivística (NOBRADE) (Conselho Nacional de Arquivos, 2006), a Lei Geral de Proteção de Dados Pessoais (LGPD), Lei de nº 13.709, de 14 de agosto de 2018⁴, a Lei de Acesso à Informação (LAI), Lei de nº 12.527, de 18 de novembro de 2011⁵, como também aos princípios arquivísticos (Gava; Flores, 2021a).

Preservar o patrimônio arquivístico digital não só é o dever das instituições arquivísticas dentro do cenário da legislação arquivística brasileira, mas antes de tudo é preservar sua memória coletiva.

2 PROCEDIMENTOS METODOLÓGICOS

O estudo consiste em uma pesquisa qualitativa, aplicada e exploratória, fundamentada na relevância prática, no aproveitamento do conhecimento prévio e na colaboração de todos os pesquisadores e envolvidos.

O projeto é dividido por funções as quais desempenham papéis específicos, sendo:

- a) Coordenador do projeto: Docente da UFES que acompanha as atividades de todas as equipes; persegue o alcance das metas estabelecidas, aprova os gastos do projeto, indica bolsistas para o projeto, presta contas do projeto junto à Fundação de Amparo à Pesquisa e Inovação do Espírito Santo (FAPES) e demais instituições envolvidas etc.
- b) Supervisor da área arquivista – APEES: Servidores do APEES que acompanham e organizam as atividades da área arquivística do projeto. Perseguir o alcance

⁴ Lei Geral de Proteção de Dados Pessoais (LGPD). Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/113709.htm.

⁵ Regula o acesso a informações previsto no inciso XXXIII do art. 5º, no inciso II do § 3º do art. 37 e no § 2º do art. 216 da Constituição Federal; altera a Lei nº 8.112, de 11 de dezembro de 1990; revoga a Lei nº 11.111, de 5 de maio de 2005, e dispositivos da Lei nº 8.159, de 8 de janeiro de 1991; e dá outras providências. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/112527.htm.

das metas estabelecidas. Entrega relatórios periódicos para o coordenador. Orienta quanto a escrita de relatórios e manuais de procedimento, desenvolvimento do Modelo de Diagnóstico de aderência ao desenvolvimento do Plano de Preservação Digital referente ao cenário de preservação dos documentos digitais do E-Docs, entre outras.

- c) Supervisor da área arquivística: Docentes da UFES que acompanham e organizam as atividades da área arquivística do projeto.
- d) Supervisor da área de TI – Prodest: Servidores do Prodest que acompanham e organizam as atividades da equipe de Analistas de Tecnologia da Informação (TI);
- e) Supervisor da área de TI - UFES: Docentes da UFES que acompanham e organizam as atividades da equipe de TI;
- f) Arquivista - APEES: Servidores do APEES que desenvolvem atividades relacionadas à área arquivística;
- g) Arquivista Especialista: Arquivistas externos especialistas que desenvolvem atividades relacionadas à área arquivística;
- h) Analista de Tecnologia de Informação: Servidores do Prodest que desenvolvem atividades relacionadas a desenvolvimento de software em colaboração com os pesquisadores;
- i) Bolsista graduado de Gestão de Projetos: Desenvolve atividades de gestão de projetos de pesquisa, sob orientação do supervisor e do coordenador do projeto.
- j) Bolsista graduado de Arquivologia: graduados em Arquivologia da UFES que desenvolvem atividades de pesquisa na área arquivística.
- k) Bolsista graduado em Ciência da Computação: graduados em Ciência da Computação da UFES que desenvolvem atividades de pesquisa na área de TI.
- l) Bolsista de graduação em Arquivologia: graduandos da UFES que desenvolvem atividades de pesquisa na área arquivística.
- m) Bolsista de graduação em Designer Gráfico: graduandos da UFES que desenvolvem atividades de pesquisa na área de design gráfico.

n) Bolsista de graduação em Ciência da Computação: graduandos da UFES que desenvolvem atividades de pesquisa na área de TI.

Para isso, estão sendo utilizados procedimentos bibliográficos e documentais, além de elementos de estudos de campo, pesquisa-ação e pesquisa participante, com a técnica de observação participante.

Como resultado, segue a prospecção de proposição de um conjunto de diretrizes conceituais, que servirão como arcabouço para um projeto estratégico nacional, por meio da realização de testes empíricos e da busca pelo desenvolvimento de soluções e oportunidades de inovação, através de implementação de práticas mais resistentes e sustentáveis para uma área de estudo.

3 RESULTADOS

Os dados relacionados referem-se aos resultados parciais do projeto desde seu início, que se deu em maio de 2024, com previsão de 24 meses. No projeto as tarefas executadas estão sendo gerenciadas por meio do Asana, uma ferramenta de gerenciamento de projetos que possibilita a colaboração em equipe. O Quadro 1 apresenta os principais resultados do projeto até o momento da escrita deste artigo:

Quadro 1 – Resultados parciais do projeto de pesquisa

Meta	Descrição	Resultados Parciais
Organização e planejamento	Compreende atividades de organização e planejamento inicial do projeto registro, tais como descrição detalhada das atividades, contratação de bolsistas, definição e configuração de software de gestão de projetos, entre outras.	Contratação de bolsistas; planejamento detalhado das atividades; desenvolvimento de Estrutura Analítica do Projeto (EAP); cadastramento das atividades no <i>software</i> de gestão Asana para acompanhamento semanal.
Nivelamentos e treinamentos	Planejamento de nivelamentos da equipe e treinamentos de assuntos diversos relevantes ao projeto.	Dois <i>workshops</i> de nivelamento para apresentação do projeto, equipe e atualização dos resultados de cada meta; treinamentos mensais para toda a equipe.
Eventos e produção acadêmica	Registro detalhado da participação em eventos científicos e programação de próximos eventos.	Participação com apresentação de trabalhos no IV Seminário Nacional de Governança Arquivística, VIII Reunião Brasileira de Ensino e Pesquisa em Arquivologia (REPARQ), X

		Congresso Nacional de Arquivologia (CNA); 8ª Semana Nacional de Arquivos, Semana do Dia do Arquivista (Prodest/Ufes).
Estudo de aderência do APEES ao RDC-Arq	Desenvolvimento de modelo de diagnóstico que possibilitará a uma instituição arquivística uma autoavaliação que verifique seu grau de aderência à normativas relativas à implementação de um RDC-Arq (Repositório Arquivístico Digital Confiável), que contempla um estudo específico do APEES.	Desenvolvimento de dois relatórios parciais: 1) Análise de documentação interna do APEES e Prodest para identificar as evidências de conformidade do APEES com as normativas relativas à implementação do RDC-Arq; 2) Desenvolvimento da parte 1 (Eixo A da Res. Nº 51 do Conarq) de um Modelo de Diagnóstico de aderência de uma instituição arquivística ao RDC-Arq.
Levantamento e estudo das principais normativas	Estudo das principais normativas estaduais, nacionais e internacionais aplicáveis à preservação digital de documentos arquivísticos, incluindo análise de requisitos e diretrizes.	Relatório sobre normativas do ES, normativas nacionais e internacionais relacionadas à preservação digital de documentos arquivísticos.
Estudo e definição de metadados	Estudo dos principais padrões de metadados nacionais e internacionais com o objetivo de definir um conjunto mínimo de metadados, aliado à análise dos metadados utilizados no sistema E-Docs por meio do mapeamento dos fluxos documentais.	Análise comparativa (<i>crosswalk</i>) entre os principais padrões nacionais e internacionais de metadados aplicáveis à gestão, preservação e acesso de documentos arquivísticos digitais. Levantamento e mapeamento dos metadados atualmente utilizados no sistema E-Docs, visando identificar lacunas e propor um conjunto mínimo de metadados alinhado às normativas vigentes.
Plano de Preservação Digital	Definir um modelo de Plano de Preservação Digital (PPD) e desenvolver um PPD para o cenário de preservação dos documentos arquivísticos digitais do E-Docs.	Definição dos elementos que devem constar na primeira versão do PPD (Introdução, Propósito, Política de Preservação Digital do Governo do Estado do Espírito Santo, Análise Contextual, Cenário de Preservação Digital, Responsabilidades, Estratégias de Preservação, Fluxo dos Usuários e Produtores, Plano de Ação, Revisão e Atualizações e Referências).
Diagnóstico do sistema E-Docs	Análise detalhada do sistema E-Docs do ponto de vista arquivístico.	Mapeamento dos fluxos documentais, identificação de pontos de melhoria e diagnóstico do E-Docs do ponto de vista da avaliação dos documentos.

Arquitetura computacional	Desenvolver uma arquitetura computacional que permita a interoperabilidade segura do E-Docs com os ambientes de preservação e acesso/difusão de documentos arquivísticos digitais, garantindo a autenticidade e acessibilidade dos documentos.	Diagnóstico computacional do sistema E-Docs; Estudo de conformidade entre a arquitetura dos sistemas de preservação e acesso com as utilizadas pelo Prodest. Instalação de ambiente de teste do Archivematica ⁶ e AtoM ⁷ .
---------------------------	--	--

Fonte: Elaboração própria.

Os resultados parciais demonstram um avanço significativo nas atividades de organização, planejamento e capacitação da equipe. Observa-se nas atividades iniciais que as metas de contratação de bolsistas, elaboração da Estrutura Analítica do Projeto (EAP) e planejamento detalhado das tarefas foram cumpridas, assim como os *workshops* de nivelamento e os treinamentos mensais, que garantiram a atualização constante da equipe.

A respeito da realização dos estudos e diagnósticos, destacam-se o levantamento e análise de normativas relacionadas à Preservação Digital, o desenvolvimento de relatórios sobre o APEES e Prodest, e a elaboração do modelo de diagnóstico de aderência ao RDC-Arq. Além disso, a análise comparativa de metadados e o levantamento dos metadados utilizados no sistema E-Docs permitiram a identificação de lacunas e a proposição de um conjunto mínimo de metadados alinhado às normativas vigentes.

Acrescenta-se que durante o desenvolvimento do levantamento de normativas observou-se a necessidade de uma ampliação na pesquisa e expandiu-se para a identificação da legislação relacionada à Preservação Digital no Brasil, e considera-se que juntas as temáticas ampliam a discussão e possibilitam um maior embasamento para o projeto.

Por fim, com relação à parte técnica do projeto, nota-se um avanço na análise detalhada do sistema E-Docs, permitindo o mapeamento de fluxos documentais e a

⁶ Archivematica é um software de código aberto para preservação digital baseado no modelo OAIS. Ele automatiza processos de submissão, normalização e armazenamento de pacotes de preservação digital, garantindo a autenticidade e acessibilidade dos documentos a longo prazo. Sistema desenvolvido pela Artefactual Systems.

⁷ Access to Memory (AtoM) é um sistema de código aberto para descrição arquivística e acesso a acervos digitais e físicos, baseado nos padrões internacionais como ISAD(G) e NOBRADE. Ele permite a organização, pesquisa e difusão de documentos arquivísticos, facilitando o acesso público a registros históricos e administrativos. Sistema desenvolvido pela Artefactual Systems.

identificação de pontos de melhoria. Aponta-se que a instalação de um ambiente teste para o Archivematica e o AtoM, juntamente com o estudo de conformidade da arquitetura computacional do E-Docs com os sistemas de preservação e acesso, estão a possibilitar uma base de conhecimento que garanta a autenticidade e acessibilidade dos documentos arquivísticos digitais.

Destacam-se os seguintes resultados previstos: a realização de um estudo de compliance para garantir a aderência do Arquivo Público do Espírito Santo à legislação arquivística, impulsionando a implementação de Repositórios Arquivísticos Digitais Confiáveis; o desenvolvimento de um Plano de Preservação Digital robusto para os documentos geridos pelo E-Docs; a elaboração de um estudo detalhado, com aplicação da abordagem MVP, que viabilize o desenvolvimento de soluções tecnológicas para a Preservação Digital Sistêmica no âmbito governamental; a qualificação técnica de profissionais das áreas arquivística e de TI; e a identificação de tendências emergentes que contribuirão para o fortalecimento da base de conhecimento técnico-científico, orientando futuras pesquisas na área.

4 CONSIDERAÇÕES FINAIS

No contexto dessa proposta de trabalho, a Preservação Digital Sistêmica é entendida como um modelo de gestão de preservação digital que deve ser seguida pelas instituições arquivísticas brasileiras, com foco não somente na obsolescência de suportes, formatos, *softwares* e *hardwares*, mas em todo o ciclo de vida do documento arquivístico digital, contemplando todos os ambientes envolvidos, que são o ambiente de produção, preservação e de acesso e difusão.

Um projeto de pesquisa que visa a Preservação Digital Sistêmica de documentos arquivísticos digitais no âmbito de todo um estado, por si só se justifica por sua relevância, pela importância da preservação do patrimônio arquivístico digital dos documentos produzidos pelos órgãos e entidades do Executivo Estadual do Governo do Estado do Espírito Santo.

Espera-se que o desenvolvimento deste promova a capacitação profissional de todos os envolvidos, por meio do fortalecimento de competências tanto na dimensão

científica quanto na tecnológica. A iniciativa visa estimular a formação crítica e técnica voltada para a compreensão e aplicação de conceitos avançados relacionados à preservação digital, com foco especial na interoperabilidade entre os ambientes de produção, preservação, acesso e difusão de documentos arquivísticos digitais.

Almeja-se a geração de conhecimento aplicado na área de PDS de documentos arquivísticos, com foco no desenvolvimento de soluções inovadoras. Como principal resultado, destaca-se a proposta de uma arquitetura computacional para a Preservação Digital Sistêmica, aplicável ao E-Docs e adaptável a outros sistemas arquivísticos em diferentes contextos geográficos.

REFERÊNCIAS

BRASIL. Lei nº 13.709, de 14 de agosto de 2018. Dispõe sobre a proteção de dados pessoais e altera a Lei nº 12.965, de 23 de abril de 2014 (Marco Civil da Internet). **Diário Oficial da União**: seção 1, Brasília, DF, 15 ago. 2018. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 08 abr. 2025.

CONSELHO INTERNACIONAL DE ARQUIVOS. **ISAD (G)**: Norma internacional de descrição arquivística. Rio de Janeiro: Arquivo Nacional, 2000. Disponível em: https://www.gov.br/conarq/pt-br/centrais-de-conteudo/publicacoes/isad_g_2001.pdf. Acesso em: 08 abr. 2025.

CONSELHO NACIONAL DE ARQUIVOS. **Diretrizes para implementação de repositórios arquivísticos digitais confiáveis (RDC-Arq)**. 2. versão. Rio de Janeiro: Arquivo Nacional, 2023. Disponível em: https://www.gov.br/conarq/pt-br/legislacao-arquivistica/resolucoes-do-conarq/Diretrizes_certificacao_rdc_arq_2023_12_12.pdf. Acesso em: 08 abr. 2025.

CONSELHO NACIONAL DE ARQUIVOS. **NOBRADE**: Norma brasileira de descrição arquivística. Rio de Janeiro: Arquivo Nacional, 2006. 124p. Disponível em: <https://www.gov.br/conarq/pt-br/centrais-de-conteudo/publicacoes/nobrade.pdf>. Acesso em: 08 abr. 2025.

CONSULTATIVE COMMITTEE FOR SPACE DATA SYSTEM (CCSDS). **Reference model for an open archival information system (OAIS)**. Washington: Magenta Book, 2024. Disponível em: <https://ccsds.org/Pubs/650x0m3.pdf>. Acesso em: 08 abr. 2025.

DORNELES, S. L.; CORRÊA, R. F.; FLORES, D. Repositórios arquivísticos digitais confiáveis (RDC-ARQ): conceitos, padrões e tecnologias. **Informação & Informação**, Londrina, v. 28, p. 179-215, 2024. Disponível em:

<https://ojs.uel.br/revistas/uel/index.php/informacao/article/view/50528>. Acesso em: 08 abr. 2025.

GAVA, T. B. S.; FLORES, D. Auditoria e certificação ao longo da cadeia de custódia digital arquivística. **Informação & Informação**, Londrina, 26, p. 424-449, 2021a. Disponível em: <https://ojs.uel.br/revistas/uel/index.php/informacao/article/view/44504>. Acesso em: 08 abr. 2025.

GAVA, T. B. S.; FLORES, D. Cadeia de custódia digital arquivística (CCDA). In: ARQUIVO, DOCUMENTO E INFORMAÇÃO EM CENÁRIOS HÍBRIDOS; SIMPÓSIO INTERNACIONAL DE ARQUIVOS, 8., 2021, São Paulo. **Anais [...]**. São Paulo: Eventus, 2021b. Disponível em: https://www.academia.edu/46918190/Cadeia_de_Cust%C3%B3dia_Digital_Arquiv%C3%ADstica_CCDA_Simp%C3%B3sio_Internacional_de_Arquivos_da_Associa%C3%A7%C3%A3o_de_Arquivistas_de_S%C3%A3o_Paulo_ARQ_SP. Acesso em: 08 abr. 2025.

GAVA, T. B. S.; FLORES, D. Repositórios arquivísticos digitais confiáveis (RDC-Arq) como plataforma de preservação digital em um ambiente de gestão arquivística. **Informação & Informação**, Londrina, v. 25, p. 74-99, 2020. Disponível em: <https://ojs.uel.br/revistas/uel/index.php/informacao/article/view/50528>. Acesso em: 08 abr. 2025.

INNARELLI, H. C. **Gestão da preservação de documentos arquivísticos digitais**: proposta de um modelo conceitual. 2015. Tese (Doutorado em Cultura e Informação) - Universidade de São Paulo, São Paulo, 2015. Disponível em: <https://www.teses.usp.br/teses/disponiveis/27/27151/tde-27052015-101628/publico/HumbertoCelesteInnarelliVC.pdf>. Acesso em: 08 abr. 2025.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. **ISO 14721:2012**. Space data and information transfer systems: open archival information system – Reference model. Geneva: ISO, 2012.



4.0 Internacional

Esta licença permite compartilhamento, remixe, adaptação e criação a partir do trabalho, mesmo para fins comerciais, desde que sejam atribuídos créditos ao(s) autor(es). Conteúdos elaborados por terceiros, citados e referenciados nesta obra não são cobertos pela licença.

Como citar este trabalho:

GAVA, Tânia Barbosa Salles; ALEIXO, Diana Vilas Boas Souto; PIGNATARO, Tatiana Canelhas. Preservação digital sistêmica: o início da construção de um modelo conceitual para o E-Docs/ES. In: WORKSHOP DE INFORMAÇÃO, DADOS E TECNOLOGIA, 8., Marília, SP. **Anais [...]**. Marília, SP: Universidade de Marília, 2025. DOI: <https://doi.org/10.22477/viii.widat.241>.



WIDaT 2025

VIII WORKSHOP DE INFORMAÇÃO,
DADOS E TECNOLOGIA

MARÍLIA • SP

Gestão da Informação do Conselho de Desenvolvimento Econômico Social Sustentável do Brasil: potencialidades e desafios da tecnologia e da Inteligência Artificial

*Information Management of the Brazilian Council for Sustainable Economic and Social
Development: potentials and challenges of technology and Artificial Intelligence*

Carla Maria Martellote Viola¹



<https://lattes.cnpq.br/3133945606177771>



<https://orcid.org/0000-0002-0895-8163>

Milton Shintaku²



<http://lattes.cnpq.br/8605833104600600>



<https://orcid.org/0000-0002-6476-4953>

Maison Roberto Mendonça Gonçalves³



<http://lattes.cnpq.br/4801149802103690>



<https://orcid.org/0000-0003-4056-6917>

Mirele Costa⁴



<http://lattes.cnpq.br/8547303047227327>



<https://orcid.org/0000-0002-1337-4672>

Resumo

Objetivo: Analisar as potencialidades e limitações do *software Omeka* e o uso da Inteligência Artificial para a gestão da informação, no Conselho de Desenvolvimento Econômico Social Sustentável do Brasil. Metodologia: Abordagens qualitativa e exploratória, estruturadas em três etapas: levantamento bibliográfico sobre gestão da informação e tecnologias emergentes no setor público; análise comparativa das versões de *software* de gestão digital - *Omeka Classic* e *Omeka S* - e de soluções de Inteligência Artificial; e elaboração de propostas de aplicação dessas tecnologias ao contexto do Conselho, visando otimizar a gestão de dados, a preservação digital e o suporte à tomada de decisões. Resultados: São elencadas dez

¹ Doutora em Ciência da Informação, Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict)/Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro, RJ, Brasil. Pesquisadora bolsista, Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), Brasília, DF, Brasil. carlaviola@ibict.br.

² Doutor em Ciência da Informação, Universidade de Brasília (UnB), Brasília, DF, Brasil. Coordenador de Tecnologias para Informação (COTEC), Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), Brasília, DF, Brasil. shintaku@ibict.br.

³ Bacharel em Biblioteconomia, Universidade Federal do Rio Grande (FURG), Rio Grande, RS, Brasil. Pesquisador bolsista, Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), Brasília, DF, Brasil. maisongoncalves@ibict.br.

⁴ Mestre em Informática pela Universidade de Brasília (UnB), Brasília, DF, Brasil. Pesquisadora bolsista, Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), Brasília, DF, Brasil. carlaviola@ibict.br. mirelecosta@ibict.br.

categorias de análise para descrever as peculiaridades de cada versão do *Omeka* e descritas nove formas de como a Inteligência Artificial pode ser relacionada ao *Omeka S*, vez que foi considerada a versão apropriada para atender ao Conselho. Considerações finais: A migração do *Omeka Classic* para o *Omeka S*, lançado em 2012, traz benefícios significativos em funcionalidade, escalabilidade e modernização, melhorando a gestão de informações, a interoperabilidade e a qualidade das coleções digitais, desde que planejada com cuidado para uma transição eficiente. Além de que, a integração da Inteligência Artificial com o *Omeka S* pode aprimorar a funcionalidade, a organização e a interação com os dados, aproveitando a flexibilidade e modernidade do *software* para gestão e publicação de coleções digitais na web.

Palavras-chave: Gestão da Informação; tecnologia; *Omeka*; Inteligência Artificial; Conselho de Desenvolvimento Econômico Social Sustentável do Brasil.

Abstract

Objective: To analyze the potential and limitations of the Omeka software and the use of Artificial Intelligence for information management at the Brazilian Council for Sustainable Social Economic Development. **Methodology:** Qualitative and exploratory approaches, structured in three stages: bibliographic survey on information management and emerging technologies in the public sector, comparative analysis of digital management software versions - Omeka Classic and Omeka S - and Artificial Intelligence solutions, and development proposals for applying these technologies to the context of the council, aiming to optimize data management, digital preservation, and support for decision-making. **Results:** Ten categories of analysis are listed to describe the peculiarities of each version of Omeka. Nine ways in which Artificial Intelligence can be related to Omeka S are described, since it was considered the appropriate version to serve the Council. **Final considerations:** Migrating from Omeka Classic to Omeka S, launched in 2012, brings significant benefits in functionality, scalability and modernization, improving information management, interoperability and the quality of digital collections, if it is carefully planned for an efficient transition. In addition, the integration of Artificial Intelligence with Omeka S can improve functionality, organization and interaction with data, taking advantage of the flexibility and modernity of the *software* for managing and publishing digital collections on the web.

Keywords: information Management; technology; *Omeka*; Artificial Intelligence; Brazilian Council for Sustainable Economic and Social Development.

1 INTRODUÇÃO

Em 2003, no primeiro governo do presidente Luiz Inácio Lula da Silva, mais conhecido como Lula, foi criado o Conselho de Desenvolvimento Econômico e Social (CDES), pelo Decreto nº 4.759, de 22 de junho de 2003, com a proposta de ser um espaço de diálogo entre governo, setor produtivo e sociedade civil, para discutir políticas públicas e estratégias de desenvolvimento sustentável (Brasil, 2003). O CDES teve um papel importante na formulação de políticas como o Plano de Aceleração do Crescimento (PAC) e o fortalecimento de políticas sociais. Com a crise política e econômica de meados da década de 2010, a atuação do Conselho foi reduzida, e ele acabou sendo descontinuado durante o governo Bolsonaro (2019-2022), pelo Decreto nº 9.759, de 11 de abril de 2019 (Brasil, 2019).

Em 2023, no terceiro mandato do Presidente Lula, o Conselho de Desenvolvimento Econômico e Social foi recriado sob o nome de Conselho de Desenvolvimento Econômico Social Sustentável (CDESS), que é popularmente conhecido como Conselhão. Essa nova versão do Conselho ampliou sua atuação para incluir temas como sustentabilidade, inclusão digital e inteligência artificial, reforçando a importância da inovação tecnológica na formulação de políticas públicas. O CDESS tem um papel estratégico na interlocução entre governo e sociedade, promovendo debates sobre desenvolvimento econômico sustentável, inclusão social e os impactos da tecnologia na economia e no mercado de trabalho (Brasil, 2023).

A recriação do CDESS em 2023 ocorre em um cenário de demandas por eficiência na gestão de dados públicos. Entre os desafios, destacam-se a desorganização de acervos digitais, a falta de integração entre bases de dados governamentais e a carência de ferramentas analíticas para embasar decisões estratégicas. Nesse contexto, a Inteligência Artificial (IA) surge como solução potencial, mas sua implementação exige superar obstáculos técnicos e éticos, como viés algorítmico e privacidade.

Dessa forma, as questões de pesquisas que se impõem são: Qual a ferramenta a ser utilizada para publicar e compartilhar coleções digitais de forma organizada para amplo acesso de seus membros, da Presidência e de toda a sociedade? Como os modelos de IA podem auxiliar na avaliação de dados, propostas, detecção de tendências e impacto social das políticas discutidas no Conselhão?

A partir desses questionamentos, este artigo tem como objetivo analisar as potencialidades e limitações do *software* Omeka e o uso da Inteligência Artificial para a gestão da informação no Conselho de Desenvolvimento Econômico Social Sustentável do Brasil.

O estudo se justifica em razão da gestão transparente e da preservação da memória do Conselhão dependerem da qualidade dos dados, das informações e dos documentos disponíveis. A informatização das atividades do Conselhão exige a criação de um ecossistema de informação que integre diferentes fontes de dados e forneça suporte analítico para os tomadores de decisão. Tecnologias e sistemas de apoio à decisão desempenham um papel fundamental nesse cenário.

2 PROCEDIMENTOS METODOLÓGICOS

A pesquisa ora proposta adotou uma abordagem qualitativa e exploratória, com o objetivo de identificar ferramentas e tecnologias que possam otimizar a gestão de dados, a organização da informação e a preservação digital no contexto do Conselho de Desenvolvimento Econômico Social Sustentável. A metodologia foi dividida em três etapas principais: levantamento bibliográfico, análise de ferramentas e tecnologias e proposta de aplicação.

A primeira etapa consistiu em um levantamento bibliográfico, com foco nos seguintes eixos temáticos: (1) a gestão da informação e a preservação digital no setor público; e (2) os usos de tecnologias emergentes como a Inteligência Artificial (IA) na administração pública. Para isso, foram consultados artigos científicos, livros, teses e documentos oficiais.

Na segunda etapa, foram analisadas técnicas da gestão da informação já aplicadas ao CDESS como proposta de melhoria na gestão de dados e na preservação digital. Constatou-se o uso pelo CDESS do *software Omeka Classic* para a gestão das coleções, por essa razão, adotou-se a comparação desse *software* com uma versão mais recente denominada *Omeka S*, uma plataforma de código aberto para a criação e gestão de coleções digitais. Conforme a documentação oficial, o *Omeka S* é uma plataforma flexível para a gestão de coleções digitais e exposições *online*, sendo adequado para projetos que exigem organização e acesso amplo a informações. Além disso, foi investigado o uso de ferramentas de IA a serem utilizadas para a análise de dados e suporte à tomada de decisões (Omeka, 2025a, 2025b).

A terceira etapa consistiu na elaboração de uma proposta de aplicação das ferramentas e tecnologias analisadas ao contexto do CDESS. Foram considerados aspectos como a integração de diferentes fontes de dados, a criação de um ecossistema de informação e a implementação de sistemas de apoio à decisão baseados em IA. A proposta foi fundamentada nas lacunas e oportunidades identificadas durante o levantamento bibliográfico e a análise de ferramentas.

2.1 Perspectivas Contemporâneas da Gestão da Informação em Confluência com a Inteligência Artificial

A evolução da tecnologia tem sido um dos principais impulsionadores da transformação digital, impactando diretamente a forma como a informação é gerenciada e utilizada pelo governo. A Inteligência Artificial (IA), como um dos pilares dessa revolução, redefine os processos de gestão da informação ao introduzir automação, aprendizado de máquina e análise preditiva para melhorar a tomada de decisão (Russell; Norvig, 2021).

A IA pode ser definida como a capacidade de sistemas computacionais realizados em tarefas que, tradicionalmente, exigiam inteligência humana, como reconhecimento de padrões, cálculo lógico e aprendizado contínuo (Mitchell, 2019). No contexto da gestão da informação, a IA possibilita um tratamento mais eficiente de grandes volumes de dados, transformando-os em conhecimento estratégico. Davenport e Prusak (1998) destacam que a informação somente se torna um recurso importante quando é organizada e interpretada dentro de um contexto específico.

Com o advento da computação em nuvem, a gestão da informação tornou-se mais dinâmica e descentralizada. Brynjolfsson e McAfee (2016) argumentam que o uso de algoritmos avançados melhora a produtividade empresarial ao reduzir ineficiências e possibilitar análises preditivas mais precisas. Essa nova abordagem reforça a necessidade de um modelo de governança de dados sólido, pautado na ética e na transparência do uso da informação (Floridi, 2015).

Além disso, a interação entre IA e gestão da informação também gera desafios. A confiabilidade dos algoritmos, a privacidade dos dados e a visão algorítmica são questões críticas que exigem regulamentação e controle. Segundo Pasquale (2016), a “caixa-preta” da IA representa um obstáculo à transparência, pois muitas decisões automatizadas não podem ser facilmente explicadas ou auditadas.

Por outro lado, a IA transforma desafios tradicionais, como alto volume de dados, falta de padronização e vulnerabilidade a erros, em oportunidades, promovendo eficiência, segurança e praticidade. Ao otimizar processos, a IA permite maior controle de produtividade, melhoria na qualidade dos serviços e resultados mais satisfatórios,

fortalecendo a competitividade e a postura estratégica das empresas (Berno; Peixe; Balsan, 2024).

Dessa forma, Canton, Miguel e Machado (2024) asseveram que as ferramentas de Gestão da Informação aliadas à Inteligência Artificial são fundamentais para a modernização e gestão dos processos, permitindo que instituições não apenas acompanhem, mas liderem a transformação digital em seus setores. A integração entre essas áreas é um campo em constante evolução, com inovações surgindo diariamente.

Nesse contexto, a IA surge como uma solução inovadora, trazendo vantagens tanto para organizações quanto para usuários, ao oferecer praticidade, conveniência e segurança. Sua aplicação na gestão documental e de dados representa uma transformação significativa e duradoura no cenário empresarial, impulsionando a excelência e o valor de entidades públicas e privadas. À medida que novas aplicações são desenvolvidas, espera-se que a IA desempenhe um papel cada vez mais central na condução de negócios de forma eficaz e sustentável (Berno; Peixe; Balsan, 2024).

Portanto, a convergência entre tecnologia, IA e gestão da informação representa um avanço significativo para a eficiência organizacional, mas exige um olhar atento às questões éticas e regulatórias. A implementação bem-sucedida dessas tecnologias deve equilibrar inovação e responsabilidade, garantindo que o uso da informação seja benéfico para todos os envolvidos.

3 RESULTADOS

Inicialmente, realizou-se uma análise sob a perspectiva sistêmica comparando as duas versões do software *Omeka*: o *Omeka Classic* (versão de 2008) e o *Omeka S* (versão de 2012). Abaixo, elenca-se as principais categorias de análise para descrever as peculiaridades de cada versão. A versão “S”, de 2012, traz uma série de benefícios significativos, especialmente em termos de funcionalidade, escalabilidade e modernização da plataforma. A migração do *Omeka Classic* para o *Omeka S* oferece vantagens que podem melhorar a qualidade da gestão das informações, a escalabilidade e a interoperabilidade de suas coleções digitais. Ao se considerar a migração, é

importante planejar cuidadosamente o processo para garantir uma transição suave e aproveitar ao máximo os benefícios da nova plataforma (Quadro 1).

Quadro 1 – Comparação entre o Omeka S e o Omeka Classic

Categoria	Omeka S	Omeka Classic
Arquitetura	Utiliza PHP 7.4 ou versões superiores, como o PHP 8, e <i>Zend Framework 3</i> , oferecendo melhor desempenho e escalabilidade.	Baseado em versões antigas de PHP e <i>Zend Framework</i> . Suportando versões atuais, mas não otimizadas para elas.
Desempenho e Segurança	Melhor desempenho e segurança, com suporte a tecnologias modernas.	Desempenho e segurança inferiores devido à base tecnológica antiga.
Integração com Padrões	Suporta JSON-LD, APIs RESTful, RDF, OAI-PMH, Schema.org	Suporta JASON, APIs RESTful e OAI-PMH.
Interoperabilidade	Suporte nativo a vocabulários controlados e <i>Linked Data</i> para interoperabilidade.	Suporte limitado a vocabulários controlados; sem suporte nativo com <i>Linked Data</i> .
Flexibilidade e Customização	Arquitetura modular, moderna e flexível. Grande integração entre os módulos.	Arquitetura fixa item - coleção - exposição. <i>Plugins</i> com pouca integração.
Escalabilidade	Permite criar e gerenciar múltiplos sites a partir de uma única instalação.	Suporta um site por instalação, sendo adequado para projetos pequenos e menos complexos.
Interface do Usuário	Interface moderna, intuitiva e limpa.	Interface mais antiga e poluída, dependendo da quantidade de <i>plugins</i> instalados.
Controle de usuários	Sistema de permissões com papéis fixos, porém com mais controle de permissões.	Sistema de permissões básico, com papéis fixos. Adequado para necessidades menos complexas.
Curva de aprendizagem	Maior curva de aprendizagem devido à estrutura modular e suporte à múltiplos sites. Requer mais configurações manuais.	Menor curva de aprendizagem devido a estrutura fixa. Requer menos configurações antes de começar a usar.
Comunidade	Comunidade ativa e com desenvolvimento de módulos e temas em crescimento constante. Suporte contínuo da comunidade.	Comunidade ativa e consolidada, com grande quantidade de instalações registradas. Suporte contínuo da comunidade.
Atualizações e suporte	Recebe atualizações regulares.	Menor incidência de atualizações.

Fonte: Elaboração própria.

A relação entre IA e o *Omeka S* pode ser explorada em vários aspectos, especialmente considerando que o *Omeka S* é uma plataforma flexível e moderna para a gestão de coleções digitais e a publicação de conteúdos na web. A IA pode ser integrada

ao *Omeka S* para melhorar a funcionalidade, a organização e a interação com os dados. A seguir são descritas 9 (nove) formas de como a IA pode ser relacionada ao *Omeka S*:

(1) Automatização de Metadados: (a) Descrição Automática de Itens - Ferramentas de IA, como modelos de processamento de linguagem natural (NLP), podem ser usadas para gerar automaticamente descrições, títulos ou *tags* para os itens digitais carregados no *Omeka S*; (b) Reconhecimento de Imagens - Algoritmos de visão computacional podem analisar imagens e gerar metadados descritivos, como identificação de objetos, cores, ou até mesmo reconhecimento facial; (c) Transcrição Automática: Para arquivos de áudio ou vídeo, a IA pode ser usada para gerar transcrições automáticas, facilitando a indexação e a busca. Nessas situações, Rodriguez, Bollen e Van de Sompel (2008) exploram a geração automática utilizando redes associativas para descrever itens digitais, enquanto Brito e Martins (2021) analisam ferramentas de geração automática e semiautomática de metadados, incluindo a transcrição automática de conteúdos audiovisuais para facilitar a indexação.

(2) Melhoria na Busca e Descoberta: (a) Busca Semântica - A IA pode ser integrada ao *Omeka S* para oferecer buscas semânticas, onde os usuários podem encontrar itens relacionados mesmo que não usem exatamente as mesmas palavras-chave; (b) Recomendações Personalizadas - Sistemas de recomendação baseados em IA podem sugerir itens ou coleções relacionadas aos usuários, com base em seu comportamento de navegação ou interesses. Em busca dessas melhorias, Mendonça e Arakaki (2024) abordam o enriquecimento semântico de metadados, melhorando a qualidade e a utilidade dos dados por meio da adição de informações contextuais e relações semânticas, o que aprimora a busca e descoberta em sistemas como o *Omeka S*. Somam-se a eles Martínez-Romero *et al.* (2010), que também destacam a aplicação de IA para gerar recomendações de metadados, o que pode ser adaptado para sugerir itens ou coleções relacionadas aos usuários com base em seu comportamento de navegação.

(3) Análise de Dados e Insights: (a) Análise de Coleções - Ferramentas de IA podem ser usadas para analisar grandes volumes de dados dentro do *Omeka S*, identificando padrões, tendências ou conexões entre itens; (b) Classificação Automática - Algoritmos de *machine learning* podem classificar automaticamente itens em categorias

ou coleções, com base em seus metadados ou conteúdo. A partir dessas necessidades, Colavizza *et al.* (2021) fornecem uma visão geral dos debates atuais e perspectivas futuras sobre o uso de IA em arquivos, em comunhão com Brito e Martins (2021), que discutem a aplicação de técnicas de aprendizado de máquina para classificar automaticamente itens em categorias ou coleções, com base em seus metadados ou conteúdo.

(4) Acessibilidade: (a) Tradução Automática - A IA pode ser usada para traduzir automaticamente metadados e conteúdos no *Omeka S*, tornando as coleções acessíveis a um público global; (b) Descrição de Imagens para Deficientes Visuais: Ferramentas de IA podem gerar descrições textuais de imagens, que podem ser lidas por leitores de tela, melhorando a acessibilidade. Nesses casos, Mendonça e Arakaki (2024) exploram o enriquecimento semântico de metadados e Martínez-Romero *et al.* (2019) discutem a geração de metadados a partir de dados visuais, o que pode ser adaptado para criar descrições textuais de imagens, melhorando a acessibilidade para deficientes visuais.

(5) Integração com *Linked Data* e Ontologias: (a) Enriquecimento de Metadados - A IA pode ser usada para vincular automaticamente itens no *Omeka S* a vocabulários controlados, ontologias ou bases de conhecimento externas, como o *Wikidata* ou *DBpedia*; (b) Identificação de Relações - Algoritmos de IA podem identificar relações entre itens ou entidades dentro das coleções, ajudando a criar conexões semânticas mais ricas. Considerando tais perspectivas, Mendonça e Arakaki (2024) abordam o enriquecimento semântico de metadados, vinculando automaticamente itens a vocabulários controlados, ontologias ou bases de conhecimento externas, como o *Wikidata* ou *DBpedia*, e Martínez-Romero *et al.* (2019) discutem a identificação de relações entre itens ou entidades dentro das coleções, ajudando a criar conexões semânticas mais ricas.

(6) Preservação Digital: (a) Detecção de Degradação - A IA pode ser usada para monitorar arquivos digitais no *Omeka S*, detectando sinais de degradação ou corrupção e sugerindo ações de preservação; (b) Restauração de Conteúdo - Em casos de arquivos danificados - imagens, áudios, vídeos, a IA pode ajudar a restaurar ou melhorar a qualidade do conteúdo. Quanto a essas questões, Arakaki *et al.* (2018) discutem o uso da Web Semântica e do padrão de metadados PREMIS na proposta do *Linked Data* para a preservação digital, incluindo a detecção de degradação de arquivos digitais. Além de que,

Duranti e Rogers (2024) exploram como a IA pode ser aplicada para preservar, gerenciar e fornecer acesso a documentos históricos e culturais, incluindo a restauração de conteúdo danificado.

(7) Interação com Usuários: (a) *Chatbots* e Assistentes Virtuais - A IA pode ser usada para criar *chatbots* ou assistentes virtuais que ajudam os usuários a navegar pelas coleções do *Omeka S*, responder perguntas ou fornecer suporte; (b) Análise de Sentimentos - Ferramentas de IA podem analisar *feedbacks* ou comentários dos usuários para entender melhor suas necessidades e melhorar a experiência na plataforma. Nessas questões, Grilo (2019) explora o conceito de assistentes virtuais baseados em IA para auxiliar os usuários, e Martínez-Romero *et al.* (2019) discutem a aplicação de IA para analisar *feedbacks* ou comentários dos usuários.

(8) Expansão de Funcionalidades com módulos: (a) O *Omeka S* é altamente extensível, e módulos podem ser desenvolvidos para integrar ferramentas de IA diretamente na plataforma. Por exemplo: Módulos para reconhecimento de imagens; Módulos para transcrição automática de áudio; e Módulos para busca semântica ou recomendação de conteúdo. Autores como Martins, Lemos e Andrade (2021) analisam o esforço tecnológico para uso e implantação de softwares para gestão de coleções digitais.

(9) Educação e Pesquisa: (a) Ferramentas de Ensino - A IA pode ser usada para criar ferramentas educacionais interativas dentro do *Omeka S*, como *quizzes* automáticos ou sistemas de aprendizado adaptativo; (b) Análise de Uso - A IA pode ajudar a analisar como os usuários interagem com as coleções, fornecendo *insights* valiosos para pesquisadores e curadores. Nessas peculiaridades, Durante e Rogers (2024) discutem o uso da IA no contexto do patrimônio documental, incluindo a criação de ferramentas educacionais interativas.

Apesar de todas essas possibilidades, não se pode negligenciar que a integração da Inteligência Artificial (IA) com o *Omeka S* apresenta desafios significativos que devem ser cuidadosamente considerados. Em primeiro lugar, a implementação exige investimentos robustos em infraestrutura tecnológica, como servidores de alto desempenho e ferramentas especializadas, além da capacitação de equipes para operar e manter esses sistemas - um obstáculo para instituições públicas com recursos limitados.

Além disso, a eficácia dos algoritmos de IA está diretamente ligada à qualidade dos dados disponíveis; metadados incompletos ou inconsistentes podem gerar resultados imprecisos ou enviesados, comprometendo a confiabilidade das análises.

Outros desafios críticos incluem questões éticas e de privacidade, como o uso responsável dos dados coletados e a garantia de transparência nos processos automatizados, exigindo a adoção de políticas claras e conformidade com regulamentações. Por fim, sistemas baseados em IA demandam manutenção contínua e atualizações regulares para adaptarem-se a novas demandas e manter sua eficiência ao longo do tempo. Esses fatores destacam a necessidade de um planejamento estratégico que equilibre inovação, viabilidade técnica e responsabilidade no uso de tecnologias emergentes.

4 CONSIDERAÇÕES FINAIS

Ao longo da história, os Conselhos de Desenvolvimento no Brasil evoluíram conforme as necessidades do Estado e da sociedade. Desde sua origem, como instrumentos de planejamento econômico, até sua configuração atual, como espaços de participação social, eles refletem as mudanças nas prioridades políticas e econômicas do país. O desafio contemporâneo do CDESS é garantir uma governança eficaz, baseada em dados, inteligência artificial e tecnologias da informação, para apoiar a formulação de políticas públicas em um contexto de transformações globais.

A gestão da informação é um pilar essencial para a transparência e eficiência na administração pública, o que justifica a necessidade de pesquisa sobre o tema. Além disso, destaca-se que a IA pode auxiliar na avaliação de dados, detecção de tendências e impacto social de políticas públicas, reforçando a relevância de explorar essas tecnologias no contexto do CDESS.

Concomitantemente, sabe-se que a preservação digital enfrenta desafios como a coleta, representação e organização de informações, o que reforça a necessidade de ferramentas robustas para garantir a integridade dos dados e sua acessibilidade. A integração da IA com o *Omeka S* representa uma oportunidade significativa para modernizar a gestão de acervos digitais em instituições públicas, promovendo maior

eficiência, acessibilidade e engajamento com o público. No entanto, é essencial abordar os desafios associados de forma estratégica e ética para garantir o sucesso e a sustentabilidade dessas iniciativas.

Dessa forma, conclui-se que a migração do *Omeka Classic* para o *Omeka S*, lançado em 2012, traz benefícios significativos em funcionalidade, escalabilidade e modernização, melhorando a gestão de informações, a interoperabilidade e a qualidade das coleções digitais, desde que planejada com cuidado para uma transição eficiente. Além de que, a integração da IA com o *Omeka S* pode aprimorar a funcionalidade, a organização e a interação com os dados, aproveitando a flexibilidade e modernidade da plataforma para a gestão e publicação de coleções digitais na web. Destaca-se que a IA tem o potencial de transformar a forma como os governos tomam decisões e gerenciam informações, reforçando a importância de uma abordagem inovadora.

A proposta ora apresentada busca integrar conhecimentos teóricos e práticos, com o objetivo de contribuir para a melhoria da gestão da informação e da transparência no CDESS. O levantamento bibliográfico forneceu o embasamento teórico necessário, enquanto a comparação das versões do software *Omeka* permitiram identificar soluções viáveis.

Por fim, a futura aplicação trará contribuições concretas para a atuação do Conselho, garantindo a preservação da memória institucional e a eficiência na tomada de decisões. Contudo, outros estudos são necessários para o aprimoramento da gestão da informação, a partir da nova versão do *software* utilizada e de novas pesquisas para adequação do uso da IA.

REFERÊNCIAS

- ARAKAKI, Felipe Augusto *et al.* Web semântica e preservação digital: o padrão de metadados PREMIS na proposta do linked data. **Informação e Tecnologia** (ITEC), João Pessoa, v. 5, n. 1, p. 141-156, 2018. Disponível em: <https://periodicos.ufpb.br/index.php/itec/article/view/38118>. Acesso em: 12 abr. 2025.
- BERNO, Adriana; PEIXE, Adriana Maria Miguel; BALSAN, Jorge. O uso da Inteligência Artificial na gestão de documentos e de dados. **P2P e Inovação**, Rio de Janeiro, v. 11, n.

1, p. 1-19, 2024. Disponível em: <https://revista.ibict.br/p2p/article/view/7068/6807>. Acesso em: 25 fev. 2025.

BRASIL. Decreto nº 4.759, de 21 de junho de 2003. Cria o Conselho de Desenvolvimento Econômico e Social (CDES), e dá outras providências. **Diário Oficial da União**: seção 1, Brasília, DF, p. 9, 23 jun. 2003. Disponível em: http://www.planalto.gov.br/ccivil_03/decreto/2003/d4759.htm. Acesso em: 25 fev. 2025.

BRASIL. Decreto nº 9.759, de 11 de abril de 2019. Extingue e estabelece diretrizes, regras e limitações para colegiados da Administração Pública Federal. **Diário Oficial da União**: seção 1, Brasília, DF, p. 5, 11 abr. 2019. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2019-2022/2019/decreto/D9759.htm. Acesso em: 25 fev. 2025.

BRASIL. Decreto nº 11.454, de 24 de março de 2023. Dispõe sobre o Conselho de Desenvolvimento Econômico Social Sustentável da Presidência da República. **Diário Oficial da União**: seção 1, Brasília, DF, p. 1, 24 mar. 2023. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2023-2026/2023/decreto/D11454.htm. Acesso em: 25 fev. 2025.

BRITO, Ana Maria Soares; MARTINS, Luciana Maria. Ferramentas de geração automática e semiautomática de metadados: uma análise crítica. **Informação e Informação**, Londrina, v. 26, n. 1, p. 1-20, 2021. Disponível em: <https://ojs.uel.br/revistas/uel/index.php/informacao/article/view/45543>. Acesso em: 25 fev. 2025.

BRYNJOLFSSON, Erik; MCAFEE, Andrew. **The second machine age**: work, progress, and prosperity in a time of brilliant technologies. Nova York: W. W. Norton and Company, 2016.

CANTON, Eduardo Helou Giraldele; MIGUEL, Renan P.; MACHADO, Andreia de Oliveira. Inteligência Artificial e Gestão da Informação. **Fatec Assis**, Assis, p. 12, 2024. Disponível em: <https://ric.cps.sp.gov.br/bitstream/123456789/23142/5/Intelig%C3%A2ncia%20artificial%20na%20gest%C3%A3o%20da%20informa%C3%A7%C3%A3o.pdf>. Acesso em: 26 fev. 2025.

COLAVIZZA, Giovanni *et al.* Archives and AI: an overview of current debates and future perspectives. **ACM Journal on Computing and Cultural Heritage (JOCCH)**, [Nova York], v. 15, n. 1, p. 1-15, 2021. Disponível em: <https://arxiv.org/abs/2105.01117>. Acesso em: 12 abr. 2025.

DAVENPORT, Thomas H.; PRUSAK, Laurence. **Working knowledge**: how organizations manage what they know. Boston: Harvard Business School Press, 1998.

DURANTI, Luciana; ROGERS, Corinne (ed.). **Artificial Intelligence and Documentary Heritage**. [S. l.]: Memory of the World (UNESCO), 2024. Disponível em: <https://interparestrustai.org/assets/public/dissemination/SCEaRNewsletterSpecialIssue2024ArtificialIntelligence.pdf>. Acesso em: 12 abr. 2025.

FLORIDI, Luciano. **The ethics of information**. Oxford: Oxford University Press, 2015.

GRILO, André. **Experiência do usuário em interfaces digitais**. Natal: SEDIS-UFRN, 2019. Disponível em: <https://repositorio.ufrn.br/handle/123456789/27011>. Acesso em: 13 abr. 2025.

MARTÍNEZ-ROMERO, Marcos *et al.* An approach for the automatic recommendation of ontologies using collaborative knowledge. *In*: CORCHADO, Emilio; WU, Xindong; OJA, Erkki (ed.). **Knowledge-Based and Intelligent Information and Engineering Systems**. Berlin: Springer, 2010. p. 74-81.

MARTÍNEZ-ROMERO, Marcos *et al.* Using association rule mining and ontologies to generate metadata recommendations from multiple biomedical databases. **Database: The Journal of Biological Databases and Curation**, [S. l.], v. 2019, p. 1-25, 10 jun. 2019. Disponível em: <https://doi.org/10.1093/database/baz059>. Acesso em: 12 abr. 2025.

MARTINS, Dalton Lopes; LEMOS, Daniela Lucas da Silva; ANDRADE, Morgana Carneiro de. Tainacan e Omeka: proposta de análise comparativa de softwares para gestão de coleções digitais a partir do esforço tecnológico para uso e implantação. **Informação e Informação**, Londrina, v. 26, n. 2, p. 569-595, abr./jun. 2021. Disponível em: <https://cip.brapci.inf.br/download/161606>. Acesso em: 13 abr. 2025.

MENDONÇA, Ana Carolina Novaes de; ARAKAKI, Ana Carolina Simionato. Produção científica sobre enriquecimento semântico de metadados. *In*: SEMINÁRIO NACIONAL DE CATALOGAÇÃO E TECNOLOGIA, 1., Brasília, 2024. **Anais [...]**. Brasília: Ibict, 2024. DOI: <http://doi.org/10.22477/i.sncat.17>. Disponível em: <https://labcotec.ibict.br/sncat/index.php/sncat/article/view/17>. Acesso em: 18 abr. 2025.

MITCHELL, Melanie. **Artificial Intelligence: a guide for thinking humans**. Nova York: Farrar, Straus and Giroux, 2019.

OMEKA. **Omeka classic user manual**. [2025a]. Disponível em: <https://omeka.org/classic/docs/>. Acesso em: 25 fev. 2025.

OMEKA. **Omeka S user manual**. [2025b]. Disponível em: <https://omeka.org/s/docs/user-manual/>. Acesso em: 25 fev. 2025.

PASQUALE, Frank. **The Black Box Society**: the secret algorithms that control money and information. Cambridge: Harvard University Press, 2016.

RODRIGUEZ, Marko A.; BOLLEN, Johan; VAN DE SOMPEL, Herbert. Automatic metadata generation using associative networks. **ACM Transactions on Information Systems (TOIS)**, [S. l.], v. 27, n. 2, p. 1-20, 2009. Disponível em: <https://arxiv.org/pdf/0807.0023>. Acesso em: 12 abr. 2025.

RUSSELL, Stuart; NORVIG, Peter. **Artificial Intelligence**: a modern approach. 4. ed. New Jersey: Pearson, 2021.



[4.0 Internacional](https://creativecommons.org/licenses/by/4.0/)

Esta licença permite compartilhamento, remixe, adaptação e criação a partir do trabalho, mesmo para fins comerciais, desde que sejam atribuídos créditos ao(s) autor(es). Conteúdos elaborados por terceiros, citados e referenciados nesta obra não são cobertos pela licença.

Como citar este trabalho:

VIOLA, Carla Maria Martellote et al. *Gestão da Informação do Conselho de Desenvolvimento Econômico Social Sustentável do Brasil: potencialidades e desafios da tecnologia e da Inteligência Artificial*. In: WORKSHOP DE INFORMAÇÃO DADOS E TECNOLOGIA, 8., 2025, Marília, SP. **Anais [...]**. Marília, SP: Universidade de Marília, 2025. DOI: <https://doi.org/10.22477/viii.widat.263>.



WIDaT 2025

VIII WORKSHOP DE INFORMAÇÃO,
DADOS E TECNOLOGIA

MARÍLIA • SP

Evolução da plataforma informacional para o Programa de Redução da Pobreza: desafios, soluções e resultados

Evolution of the information platform for the Poverty Reduction Program: challenges, solutions and results

Tainá Regly¹



<http://lattes.cnpq.br/5572586959935763>



<https://orcid.org/0000-0002-5127-7013>

Joyce Paola Teixeira Sobrinho²



<http://lattes.cnpq.br/1599646488165006>



<https://orcid.org/0009-0001-0598-9298>

Rebeca dos Santos de Moura³



<http://lattes.cnpq.br/8677193043257356>



<https://orcid.org/0000-0002-7685-8826>

Lucas Rodrigues Costa⁴



<http://lattes.cnpq.br/3133273170328412>



<https://orcid.org/0000-0002-0973-4866>

Diego José Macedo⁵



<http://lattes.cnpq.br/2205539000237712>



<https://orcid.org/0000-0002-5696-0639>

Resumo

A plataforma Acredita, evolução dos sistemas Progredir e SISPRP, foi desenvolvida para aprimorar a inclusão produtiva de beneficiários do Cadastro Único por meio do acesso facilitado a emprego, qualificação

¹ Doutora em Ciência da Informação, Universidade Federal do Rio de Janeiro em convênio com o Instituto Brasileiro de Informação em Ciência e Tecnologia (UFRJ-IBICT), Rio de Janeiro, RJ, Brasil. Professora Adjunta, Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro, RJ, Brasil. Pesquisadora, Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), Brasília, DF, Brasil. tainasouza@ibict.br.

² Graduação em Psicologia, Instituto de Educação Superior de Brasília (IESB), Brasília, DF, Brasil. Pesquisadora, Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), Brasília, DF, Brasil. joycesobrinho@ibict.br.

³ Mestre em Engenharia de Sistemas Eletrônicos e de Automação, Universidade de Brasília (UnB), Brasília, DF, Brasil. Desenvolvedora e assistente de pesquisa, Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), Brasília, DF, Brasil. rebecamoura@ibict.br.

⁴ Doutor em Ciência da Computação pela Universidade de Brasília (UnB), Brasília, DF, Brasil. Assistente de pesquisa, Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), Brasília, DF, Brasil. lucasrodrigues@ibict.br.

⁵ Mestre em Ciência da Informação pela Universidade de Brasília (UnB), Brasília, DF, Brasil. Tecnologista, Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), Brasília, DF, Brasil. diegomacedo@ibict.br.

profissional e empreendedorismo. Este estudo apresenta os desafios enfrentados na implementação da plataforma, as soluções adotadas e os impactos gerados. A metodologia baseou-se na análise documental e observação dos processos de desenvolvimento, com foco na migração de dados, reestruturação da arquitetura e integração com bases governamentais. Foram superados obstáculos como a compatibilidade de banco de dados, adaptação ao novo padrão visual do MDS e validação das regras de negócio. As soluções implementadas resultaram em uma plataforma mais eficiente e acessível, com melhorias na interface, adoção de APIs para interoperabilidade e otimização dos processos de homologação. Os resultados indicam avanços na usabilidade, integração e confiabilidade do sistema, tornando-o um instrumento estratégico para gestores públicos e beneficiários. A modernização do sistema possibilitou maior eficiência na gestão de políticas públicas, consolidando a plataforma como uma ferramenta essencial para a democratização do acesso a oportunidades.

Palavras-chave: inclusão produtiva; interoperabilidade; banco de dados; acessibilidade digital; políticas públicas.

Abstract

The Acredita platform, an evolution of the Progredir and SISPRP systems, was developed to improve the productive inclusion of Unified Registry beneficiaries through easier access to employment, professional qualifications and entrepreneurship. This study presents the challenges faced in implementing the platform, the solutions adopted and the impacts generated. The methodology was based on document analysis and observation of the development processes, with a focus on data migration, restructuring the architecture and integration with government databases. Obstacles such as database compatibility, adaptation to the new MDS visual standard and validation of business rules were overcome. The solutions implemented resulted in a more efficient and accessible platform, with improvements to the interface, adoption of APIs for interoperability and optimization of homologation processes. The results indicate advances in the usability, integration and reliability of the system, making it a strategic tool for public managers and beneficiaries. The modernization of the system has enabled greater efficiency in the management of public policies, consolidating the platform as an essential tool for democratizing access to opportunities.

Keywords: productive inclusion; interoperability; database; digital accessibility; public policies.

1 INTRODUÇÃO

A inclusão produtiva da população em situação de vulnerabilidade social é um dos grandes desafios das políticas públicas contemporâneas. Desde a sua criação, em 2017, o Plano Progredir tem sido um instrumento fundamental para promover oportunidades de emprego, qualificação profissional e incentivo ao empreendedorismo, visando à autonomia econômica dos cidadãos inscritos no Cadastro Único (CadÚnico). Instituído pelo Decreto nº 9.160, de 26 de setembro de 2017, o programa constitui um conjunto articulado de ações voltadas à intermediação de mão de obra, oferta de capacitação e acesso ao microcrédito (Brasil, 2017).

Com a evolução do projeto e mudanças na gestão governamental, a plataforma informacional do programa passou por transformações significativas. Inicialmente denominada Plataforma Progredir, a ferramenta foi reformulada para se tornar o Sistema

do Programa de Redução da Pobreza (SISPRP), consolidando-se como um ambiente digital voltado à inclusão produtiva. Mais recentemente, em um novo ciclo de reformulação e, a plataforma foi renomeada como Acredita para atender à nova legislação vigente, estabelecida pela Lei nº 14.995 (Brasil, 2024), incorporando diretrizes para melhorar o seu alcance e atender de maneira satisfatória as necessidades dos usuários.

A atuação do programa sempre esteve estruturada em três eixos principais:

- a) **Empreendedorismo**, promovendo acesso a microcrédito e suporte para negócios individuais e coletivos;
- b) **Intermediação de mão de obra**, conectando trabalhadores a oportunidades de emprego;
- c) **Qualificação profissional**, oferecendo cursos presenciais e a distância para capacitação e aprimoramento de habilidades.

Os objetivos do programa permanecem alinhados com sua missão inicial, buscando ampliar o acesso ao mundo do trabalho por meio da criação de condições que melhorem a empregabilidade e a renda dos beneficiários. Também se destaca o estímulo a ações municipais e estaduais de inclusão produtiva, fortalecendo parcerias locais para ampliar o impacto do programa. Por fim, há o compromisso de engajar órgãos públicos e entidades privadas na promoção da inclusão produtiva do público-alvo, garantindo maior alcance e efetividade das iniciativas implementadas.

Dado o caráter estratégico da plataforma, a coleta, organização e disponibilização de dados são aspectos essenciais para o seu funcionamento. A gestão eficiente dessas informações permite acompanhar a demanda por inclusão produtiva, identificar tendências do mercado de trabalho e mapear as necessidades de qualificação regional. Nesse contexto, a plataforma Acredita desempenha um papel central na oferta de serviços digitais integrados, facilitando o acesso dos cidadãos às oportunidades disponíveis.

O desenvolvimento e aprimoramento contínuo da plataforma trouxeram desafios significativos, como a necessidade de interoperabilidade entre bases de dados governamentais, a adequação a novas regras de negócio e a construção de uma interface acessível e intuitiva para diferentes perfis de usuários. Na comunicação "Programa

Redução da Pobreza: sistema de informação para apoiar o emprego, o empreendedorismo e a educação das pessoas inscritas no Cadastro Único", publicado por Pérez *et al.* (2024), foram discutidos os fundamentos metodológicos e técnicos para a concepção inicial da plataforma, bem como sua relevância para as políticas públicas de inclusão produtiva.

Dando continuidade a essa análise, este artigo tem como objetivo apresentar os desafios enfrentados, as soluções adotadas e os avanços tecnológicos realizados na evolução da plataforma. Serão discutidos os principais obstáculos superados no desenvolvimento do sistema, incluindo questões técnicas, de usabilidade e de interoperabilidade. Além disso, serão detalhadas as melhorias implementadas e os impactos esperados para a efetivação da plataforma como ferramenta de suporte à inclusão produtiva e ao desenvolvimento econômico dos beneficiários do programa.

2 PROCEDIMENTOS METODOLÓGICOS

O presente estudo adota uma abordagem qualitativa e descritiva, baseada na análise documental e na observação dos processos de desenvolvimento e implementação da plataforma informacional do Sistema Acredita. O método utilizado segue um modelo de pesquisa aplicada, com foco na identificação de desafios, proposição de soluções e avaliação dos avanços tecnológicos na evolução da plataforma.

A metodologia foi organizada em três etapas principais:

- a) **Levantamento e análise documental:** Esta etapa consistiu no exame detalhado de documentos institucionais, incluindo relatórios técnicos, planos de trabalho e registros de acompanhamento do projeto. Foram analisados dados referentes ao planejamento, implementação e evolução da plataforma, considerando a transição entre as diferentes fases do sistema, desde o Progredir, passando pelo SISPRP, até sua consolidação como Acredita. O levantamento documental possibilitou a compreensão dos marcos regulatórios, das exigências tecnológicas e dos desafios operacionais enfrentados ao longo do processo.

- b) **Observação e análise do desenvolvimento técnico:** Com o objetivo de compreender os desafios práticos enfrentados pela equipe de desenvolvimento, foram analisados os registros de progresso da implementação do sistema, incluindo relatórios de atraso, mudanças na arquitetura tecnológica e ajustes na modelagem de dados. A análise centrou-se nos aspectos de interoperabilidade informacional, integração com bases de dados governamentais (como CadÚnico), padronização da identidade visual e adequação a requisitos de segurança e acessibilidade.
- c) **Avaliação dos impactos e resultados:** A última etapa consistiu na identificação dos avanços obtidos com a implementação das soluções adotadas. Para isso, foram considerados indicadores de desempenho técnico, usabilidade e impacto na inclusão produtiva, conforme definidos nos relatórios do projeto. A avaliação envolveu métricas como o tempo de resposta do sistema, taxa de adoção da plataforma, engajamento dos usuários e qualidade da integração com serviços parceiros.

A pesquisa utilizou metodologias ágeis, como *Scrum*⁶ e Design Centrado no Usuário (Clark, 2011; Schwaber; Sutherland, 2020), para compreender as iterações do desenvolvimento e os ajustes realizados com base no *feedback* dos *stakeholders*⁷. O estudo também se apoia na Ciência da Informação, explorando como a organização e o fluxo de dados podem otimizar a gestão de políticas públicas voltadas à inclusão produtiva.

Com essa abordagem, busca-se fornecer uma visão abrangente dos desafios e avanços da plataforma Acredita, consolidando um panorama das transformações ocorridas desde sua concepção até o estágio atual de desenvolvimento e implementação.

⁶ Metodologia ágil usada para gerenciar projetos complexos, especialmente em desenvolvimento de software. Disponível em: <https://www.scrum.org/>. Acesso em: 12 abr. 2025.

⁷ Pessoas ou grupos com interesse direto ou indireto em um projeto, como usuários, patrocinadores ou gestores.

3 O DESENVOLVIMENTO DA PLATAFORMA

3.1 Desafios Enfrentados

O desenvolvimento da plataforma Acredita envolveu desafios técnicos e operacionais complexos. Entre os principais obstáculos enfrentados, destacam-se problemas de compatibilidade e migração de banco de dados, adaptação ao novo padrão visual, dificuldades no processo de homologação e integração das tabelas do banco de dados, conforme definido pela equipe de tecnologia da informação do Ministério do Desenvolvimento e Assistência Social, Família e Combate à Fome (MDS).

A migração de dados do antigo sistema Progridir representou um grande desafio devido à incompatibilidade do modelo de dados legado com a nova arquitetura baseada em *PostgreSQL* 15⁸. A ferramenta utilizada para modelagem, *PowerDesigner* 16.3⁹, não oferecia suporte total à versão adotada pelo MDS, exigindo ajustes manuais na conversão dos scripts de banco de dados. Além disso, o processo de dump e importação de dados gerou inconsistências nos registros, exigindo esforços adicionais para validação das tabelas e normalização das informações.

Durante essa transição, foi necessário corrigir falhas que afetavam a integridade dos registros, além de revisar relações entre tabelas e ajustar chaves primárias. Essas correções demandaram múltiplos ciclos de homologação, pois qualquer erro poderia comprometer a confiabilidade da plataforma e impactar os beneficiários.

Outro desafio crítico foi a adequação da interface da plataforma ao novo *Design System*¹⁰ do MDS. A equipe de desenvolvimento já havia estruturado a interface utilizando um modelo anterior, mas a atualização das diretrizes exigiu modificações em mais de 150 telas. O redesenho incluiu ajustes em ícones, cores, botões e layouts para garantir conformidade com a identidade visual exigida.

⁸ Versão do sistema gerenciador de banco de dados relacional open source, conhecido por sua robustez e suporte a dados complexos. Disponível em: <https://www.postgresql.org/about/news/postgresql-15-released-2510/>. Acesso em: 12 abr. 2025.

⁹ Ferramenta para modelagem de dados, arquitetura corporativa e design de banco de dados.

¹⁰ Conjunto de padrões visuais e de usabilidade para serviços digitais do governo brasileiro. Disponível em: <https://www.gov.br/ds>. Acesso em: 12 abr. 2025.

Essas mudanças impactaram diretamente o tempo de desenvolvimento, pois foi necessário revisar os estilos CSS¹¹, adaptar os componentes gráficos e revalidar toda a interface. Além disso, a acessibilidade foi aprimorada com a implementação de modo de alto contraste, suporte a leitores de tela e melhorias na navegação via teclado, garantindo que a plataforma atendesse às diretrizes do WCAG 2.1¹².

Uma complicação adicional foi a requisição, por parte do ministério, para a substituição de diversas bibliotecas utilizadas no desenvolvimento do *frontend* em React. Especificamente, foi solicitada a remoção da biblioteca *Formik*¹³, que desempenhava um papel central no sistema, sendo amplamente utilizada em fluxos que envolviam o preenchimento de formulários. Desde o login do usuário até o cadastro de empresas parceiras, cursos, vagas e outros elementos do sistema, *Formik* era a base para a manipulação e validação de dados. Essa refatoração representou um impacto substancial, pois exigiu a reescrita de partes críticas da aplicação, aumentando a complexidade e o tempo necessário para garantir que os novos métodos adotados mantivessem a integridade e a funcionalidade esperada pelo sistema.

A homologação das tabelas do banco de dados também se revelou um processo desafiador. O alinhamento entre as versões desenvolvidas pelo Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict) e os padrões exigidos pelo MDS resultou em atrasos na homologação final, uma vez que ajustes eram frequentemente necessários. Em determinados momentos, algumas tabelas essenciais foram consideradas não conformes, exigindo reformulações antes da liberação definitiva do sistema.

Além disso, modificações na estrutura do banco de dados impactaram diretamente o desenvolvimento da plataforma, afetando tanto o *backend*, que precisou ajustar os *endpoints* da API¹⁴ e revisar o código, quanto o *frontend*¹⁵, que teve que rever

¹¹ Cascading Style Sheets (CSS). Linguagem de estilo usada para definir a aparência de elementos em páginas web (cores, fontes, layout etc.). Disponível em: <https://developer.mozilla.org/pt-BR/docs/Web/CSS>. Acesso em: 12 abr. 2025.

¹² Web Content Accessibility Guidelines (WCAG 2.1). Conjunto de diretrizes internacionais para tornar o conteúdo web mais acessível a pessoas com deficiência. Disponível em: <https://www.w3.org/TR/WCAG21/>. Acesso em: 12 abr. 2025.

¹³ Biblioteca JavaScript para facilitar a criação e validação de formulários em aplicações React. Disponível em: <https://formik.org/>. Acesso em: 12 abr. 2025.

¹⁴ Pontos de acesso em uma API que permitem a comunicação entre sistemas, enviando ou recebendo dados.

¹⁵ Parte visual e interativa de um site ou aplicativo com a qual o usuário interage diretamente.

o uso dos *endpoints* modificados para refletir a nova organização dos dados. Esse retrabalho aumentou o tempo necessário para concluir a integração das funcionalidades, dificultando a estabilização da plataforma.

Os desafios enfrentados na construção da Acredita demonstram a complexidade de desenvolver um sistema governamental robusto, que exige compatibilidade com novas tecnologias, adequação a padrões institucionais e validação rigorosa das informações. A superação dessas dificuldades foi essencial para tornar a plataforma mais estável, eficiente e alinhada às diretrizes do Programa de Redução da Pobreza, garantindo uma solução integrada e confiável para os usuários.

3.2 Soluções Adotadas e Inovações Implementadas

Para superar os desafios na implementação da plataforma Acredita, diversas estratégias foram adotadas, focando na reestruturação da arquitetura, integração eficiente de dados e aprimoramento da experiência do usuário. As principais soluções envolveram a reformulação tecnológica da plataforma, a adoção do *PowerDesigner* e *PostgreSQL* para modelagem e migração de dados, o redesenho da interface para alinhamento com o novo *Design System* do governo, a implementação de um fluxo de login híbrido e o aprimoramento do processo de homologação.

A necessidade de maior escalabilidade e interoperabilidade levou à adoção de uma arquitetura modular baseada em microsserviços, garantindo maior flexibilidade e facilitando a integração com bases de dados governamentais, como o Cadastro Único (CadÚnico). Além disso, *APIs RESTful*¹⁶ foram incorporadas para permitir comunicação eficiente entre módulos da plataforma e outros sistemas, garantindo a integridade e o compartilhamento seguro de informações.

A modelagem de dados e a migração da antiga base do Progredir para a Acredita exigiram o uso do *PowerDesigner* para criação de um modelo lógico e físico do banco de dados. No entanto, incompatibilidades com o *PostgreSQL* 15, adotado pelo MDS, demandaram ajustes manuais nos scripts SQL. Para garantir a integridade dos dados na

¹⁶ Interfaces que seguem o estilo REST para permitir a comunicação entre sistemas via requisições HTTP. Disponível em: <https://restfulapi.net/>.

transição, foi utilizada a abordagem *Extract, Transform, Load* (ETL), assegurando a correta extração, transformação e carga das informações na nova base de dados.

O redesenho da interface foi necessário para alinhar o sistema ao novo *Design System* do MDS, resultando na reformulação de mais de 150 telas. Para garantir consistência, foram adotadas bibliotecas de componentes reutilizáveis, substituindo o *Font Awesome*¹⁷ pela biblioteca de ícones do Google, a *Material Symbols*¹⁸, o que reduziu o tempo de carregamento das páginas. A acessibilidade também foi aprimorada, com implementação de modo de alto contraste, suporte a leitores de tela e navegação por teclado, garantindo conformidade com as diretrizes do WCAG 2.1.

Para remover a biblioteca *Formik* do sistema, ela foi substituída pela *React Hook Form* (RHF)¹⁹, biblioteca aprovada pelo MDS. Com isso, cerca de 20 formulários foram refatorados, impactando o dobro de páginas, já que cada um é usado para submissão, visualização e edição de dados. A migração exigiu a reestruturação do envio de informações para o backend, além de ajustes no design para atender aos requisitos do RHF. Apesar dos desafios, a transição foi concluída com sucesso, garantindo a total adesão ao RHF sem comprometer o funcionamento do sistema.

A autenticação dos usuários também passou por reformulações significativas. Inicialmente prevista apenas pelo Gov.br, a plataforma precisou adotar um modelo híbrido, contemplando três métodos de login:

- a) Trabalhadores e beneficiários: via gov.br.
- b) Administradores: via LDAP (sistema interno do MDS).
- c) Parceiros e entidades privadas: via e-mail e senha próprios.

Essa mudança exigiu revisão no *backend* e ajustes na estrutura do banco de dados, garantindo a vinculação correta entre os diferentes métodos de autenticação e os perfis de usuários. Além disso, a conformidade com a LGPD foi um ponto crítico, exigindo

¹⁷ Biblioteca de ícones vetoriais e logotipos amplamente usada em interfaces web. Disponível em: <https://fontawesome.com/>.

¹⁸ Conjunto de ícones criado pelo Google com base na linguagem de design Material. Disponível em: <https://fonts.google.com/icons>.

¹⁹ Biblioteca leve para lidar com formulários em React, com foco em desempenho e facilidade de uso. Disponível em: <https://react-hook-form.com/>.

implementação de mecanismos avançados de segurança e rastreabilidade para garantir a privacidade e proteção dos dados.

A homologação do sistema foi um processo complexo, exigindo múltiplos ciclos de validação para assegurar que as funcionalidades e regras de negócio estivessem corretamente implementadas. Foram criados ambientes de teste separados, permitindo que a equipe técnica do MDS realizasse validações sem impactar o ambiente de produção. Além disso, foram adotados Testes de Aceitação Baseados em Comportamento (BDD), garantindo maior precisão na validação das regras de negócio.

A homologação das tabelas do banco de dados também precisou de ajustes contínuos, pois algumas estruturas inicialmente aprovadas passaram por reformulações. Além disso, foram realizados testes de usabilidade com usuários reais, permitindo identificar pontos de melhoria na interface e navegação da plataforma.

As soluções implementadas na Acredita garantiram maior eficiência, segurança e acessibilidade, tornando a plataforma mais robusta e integrada. A modernização da arquitetura, do banco de dados e da interface consolidou seu papel como ferramenta essencial para a inclusão produtiva e o desenvolvimento social, otimizando a experiência dos usuários e a gestão das políticas públicas.

4 RESULTADOS E IMPACTOS DA NOVA ESTRUTURA DA PLATAFORMA

A evolução da plataforma Acredita, anteriormente conhecida como Progridir e SISPRP, trouxe avanços significativos em usabilidade, integração de dados e impacto nas políticas públicas de inclusão produtiva. A reformulação da arquitetura do sistema e a adoção de novos padrões de autenticação e acessibilidade tornaram a plataforma mais eficiente, segura e acessível para diferentes perfis de usuários.

Entre as principais melhorias, destaca-se a adoção de padrões modernos de design e acessibilidade, facilitando a navegação e tornando a experiência do usuário mais intuitiva. O redesenho da interface, alinhado ao novo *Design System* do MDS, padronizou os componentes gráficos e reduziu a curva de aprendizado. Além disso, foram incorporados recursos de acessibilidade digital, como modo de alto contraste, suporte a leitores de tela e navegação por teclado, permitindo que usuários com diferentes

necessidades possam acessar os serviços de forma equitativa. A substituição do *Formik* pelo *React Hook Form* também trouxe benefícios, tornando a manipulação de formulários mais eficiente e alinhada às melhores práticas de desenvolvimento.

A reformulação do modelo de dados também foi essencial para garantir maior eficiência e integração com sistemas governamentais. A adoção do *PostgreSQL 15* como SGBD possibilitou melhorias no desempenho da plataforma, facilitando a sincronização com bases como o Cadastro Único (CadÚnico). Entre os avanços observados, destacam-se a reestruturação do banco de dados conforme os padrões do MDS, otimização da indexação de tabelas e uso de *APIs RESTful* para melhorar a comunicação entre sistemas. Como resultado, a plataforma passou a operar de forma mais estável e rápida, minimizando erros e reduzindo tempos de resposta.

As mudanças trarão impactos diretos tanto para usuários do Cadastro Único quanto para gestores de políticas públicas. Para os usuários, a nova plataforma deverá proporcionar maior facilidade no acesso a oportunidades de emprego e capacitação, aprimorando filtros de busca e recomendação de vagas com base no perfil socioeconômico. A interface mais intuitiva e a integração segura com o Gov.br simplificaram o acesso ao sistema e reduziram barreiras na navegação.

Para os gestores públicos, a reformulação trará benefícios como maior precisão na análise da demanda e oferta de qualificação profissional, facilitando o planejamento de políticas de inclusão produtiva. A extração de relatórios e indicadores foi otimizada, permitindo um monitoramento mais ágil e embasado em dados confiáveis. Com um sistema mais eficiente, foi possível reduzir o tempo de processamento das informações e tomar decisões estratégicas com maior rapidez e assertividade.

A adoção de um novo modelo de governança e monitoramento possibilitou a definição de indicadores de desempenho, permitindo a avaliação contínua da eficiência e impacto da plataforma. Alguns dos principais indicadores incluem:

- a) **Desempenho Técnico:** tempo médio de resposta, disponibilidade do sistema e volume de acessos simultâneos.
- b) **Usabilidade e Acessibilidade:** taxa de adoção da plataforma, tempo médio para completar processos e taxa de acessibilidade digital.

- c) **Impacto na Inclusão Produtiva:** número de usuários beneficiados, percentual de trabalhadores empregados após o uso da plataforma e quantidade de microcréditos concedidos.

Esses indicadores garantem que ajustes contínuos possam ser realizados para otimizar o funcionamento da plataforma.

Com essas mudanças, a Acredita se consolidou como um instrumento estratégico para a inclusão produtiva, garantindo maior eficiência na gestão de políticas públicas e promovendo um mercado de trabalho mais acessível e equitativo. A plataforma agora está preparada para evoluir continuamente, ampliando seu escopo e incorporando novas funcionalidades conforme as demandas sociais forem identificadas.

5 CONSIDERAÇÕES FINAIS

A modernização da plataforma Acredita representa um avanço significativo na inclusão produtiva, ampliando o acesso ao emprego, à qualificação profissional e ao empreendedorismo. A reformulação da arquitetura e o uso de tecnologias avançadas tornaram a plataforma mais eficiente, acessível e integrada, fortalecendo a formulação de políticas públicas baseadas em dados.

O desenvolvimento trouxe desafios como a migração de banco de dados, a integração com sistemas governamentais, a necessidade de adaptação ao novo padrão visual do MDS e a implementação de um fluxo de login híbrido. Esses obstáculos exigiram soluções estratégicas e alinhamento contínuo entre equipes técnicas e gestores. Como resultado, a plataforma oferece uma experiência mais intuitiva e melhora a interoperabilidade com bases como o CadÚnico.

A principal lição aprendida foi a importância da flexibilidade e adaptação contínua em sistemas governamentais. Ajustes na arquitetura e na interface demonstraram a necessidade de uma abordagem iterativa e colaborativa para garantir melhorias progressivas.

Para o futuro, a plataforma deve expandir sua interoperabilidade, integrar novos serviços e aprimorar funcionalidades, como o acompanhamento personalizado dos beneficiários. Tecnologias emergentes, como inteligência artificial e análise preditiva,

serão fundamentais para aprimorar a experiência do usuário. Com esses avanços, a plataforma Acredita se consolida como uma ferramenta essencial para a inclusão produtiva, contribuindo para um mercado de trabalho mais acessível e equitativo.

REFERÊNCIAS

BRASIL. Decreto nº 9.160, de 26 de setembro de 2017. Institui o Plano Progridir. **Diário Oficial da União**: seção 1, Brasília, DF, n. 186, p. 3, 27 set. 2017.

BRASIL. Lei nº 14.995 de 10 de outubro de 2024. Institui o Programa Acredita no Primeiro Passo e o Programa de Mobilização de Capital Privado Externo e Proteção Cambial. **Diário Oficial da União**: seção 1, Brasília, DF, n. 196, 11 out. 2024. Disponível em: <https://legislacao.presidencia.gov.br/atos/?tipo=LEI&numero=14995&ano=2024&ato=587kXVE9ENZpWT72d>. Acesso em: 10 mar. 2025.

CLARK, Kevin A. Human interaction and collaborative innovation. In: KUROSU, Masaaki (ed.). **Human centered design**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.

PÉREZ, Lisandra Guerrero *et al.* Programa Redução da Pobreza: sistema de informação para apoiar o emprego, o empreendedorismo e a educação das pessoas inscritas no Cadastro Único. In: WORKSHOP DE INFORMAÇÃO, DADOS E TECNOLOGIA, 7., 2024, Porto Velho. **Anais [...]**. Porto Velho: UNIR, 2024. Disponível em: <https://doi.org/10.22477/vii.widat.181>. Acesso em: 4 maio 2025.

SCHWABER, Ken; SUTHERLAND, Jeff. **The Scrum Guide**: the definitive guide to Scrum: the rules of the game. [S. l.]: [s. n.], 2020. Disponível em: <https://scrumguides.org/docs/scrumguide/v2020/2020-Scrum-Guide-US.pdf>. Acesso em: 17 fev. 2024.



[4.0 Internacional](https://creativecommons.org/licenses/by/4.0/)

Esta licença permite compartilhamento, remixe, adaptação e criação a partir do trabalho, mesmo para fins comerciais, desde que sejam atribuídos créditos ao(s) autor(es). Conteúdos elaborados por terceiros, citados e referenciados nesta obra não são cobertos pela licença.

Como citar este trabalho:

REGLY, Tainá *et al.* Evolução da plataforma informacional para o Programa de Redução da Pobreza: desafios, soluções e resultados. In: WORKSHOP DE INFORMAÇÃO DADOS E TECNOLOGIA, 8., 2025, Marília, SP. **Anais [...]**. Marília, SP: Universidade de Marília, 2025. DOI: <https://doi.org/10.22477/viii.widat.259>.



WIDaT 2025

VIII WORKSHOP DE INFORMAÇÃO,
DADOS E TECNOLOGIA

MARÍLIA • SP

Mapeamento geoespacial de áreas com potenciais problemas no acesso às unidades básicas de saúde no município de Alegre/ES

Geospatial mapping of areas with potential problems in accessing primary health care units in the municipality of Alegre/ES

Henrique Monteiro Cristovão¹



<http://lattes.cnpq.br/5035919384923489>



<https://orcid.org/0000-0003-2011-7022>

Alexsander Barbosa Moura²



<http://lattes.cnpq.br/4124477239782330>



<https://orcid.org/0009-0007-7073-5466>

Renan Carlos da Silva³



<http://lattes.cnpq.br/8441187306347073>



<https://orcid.org/0009-0004-1088-1320>

Ágata Leite Cristovão⁴



<http://lattes.cnpq.br/9101346502880308>



<https://orcid.org/0009-0002-1930-658X>

Camile Pereira de Abreu⁵



<http://lattes.cnpq.br/6218853669033686>



<https://orcid.org/0009-0006-0546-8421>

Resumo

Municípios brasileiros, em geral, enfrentam o desafio de garantir o acesso à saúde aos seus cidadãos, especialmente quando consideradas a distribuição populacional e a complexidade das rotas viárias até as Unidades Básicas de Saúde (UBS). O objetivo geral da pesquisa é identificar áreas do município de Alegre/ES que possam estar com problemas no acesso à saúde em função da distribuição geográfica das UBS, considerando-se a densidade populacional e a distância roteada das vias para acesso à UBS mais próxima. A pesquisa é de natureza aplicada, de abordagem qualiquantitativa, de caráter exploratório e descritivo, envolvendo a identificação e integração de diversas bases de dados abertas sobre dados populacionais, vias rodoviárias, geolocalização das UBS do município, e uso de sistema de informação geográfica. Como resultado, destaca-se a identificação das

1 Doutor em Ciência da Informação, Universidade de Brasília (UnB), Brasília, DF, Brasil. Professor, Universidade Federal do Espírito Santo (UFES), Vitória, ES, Brasil. henrique.cristovao@ufes.br.

2 Graduando em Engenharia de Computação, Universidade Federal do Espírito Santo (UFES), Vitória, ES, Brasil. als01bm@gmail.com.

3 Licenciatura em Letras - Língua Portuguesa e Literaturas de Língua Portuguesa, Universidade Federal do Espírito Santo (UFES), Vitória, ES, Brasil. Graduando em Estatística, Universidade Federal do Espírito Santo (UFES), Vitória, ES, Brasil. renan.silva.16@edu.ufes.br.

4 Graduanda em Estatística, Universidade Federal do Espírito Santo (UFES), Vitória, ES, Brasil. agataflor@gmail.com.

5 Graduação em Psicologia, Centro Universitário Salesiano (UniSales), Vitória, ES, Brasil. Graduanda em Artes Visuais, Universidade Federal do Espírito Santo (UFES), Vitória, ES, Brasil. camileabreudg@gmail.com.

áreas do município de Alegre com potenciais problemas no acesso à saúde alcançada por intermédio de uma aplicação web interativa implementada em Python que apresenta um mapa de calor, com granularidade de 400m, em função do índice de acesso à saúde calculado. A aplicação pode auxiliar no planejamento e gestão da saúde, identificando áreas críticas para melhor alocação de recursos e redução de desigualdades. Também pode apoiar o planejamento futuro, prevenindo problemas em regiões de crescimento rápido e fortalecendo a transparência na gestão pública. Como continuidade da pesquisa, a integração do histórico de atendimentos e CIDs permitiria um mapeamento mais preciso da incidência de doenças, auxiliando no combate de enfermidades como, por exemplo, as arboviroses.

Palavras-chave: mapeamento geoespacial; sistemas de informação geográfica; densidade populacional; unidades básicas de saúde; saúde pública.

Abstract

Brazilian municipalities, in general, face the challenge of ensuring access to health care for their citizens, especially when considering the population distribution and the complexity of road routes to Basic Health Units (UBS). The general objective of the research is to identify areas of the municipality of Alegre/ES that may have problems in accessing health care due to the geographic distribution of UBS, considering the population density and the routed distance of the roads to access the nearest UBS. The research is of an applied nature, with a qualitative-quantitative approach, of an exploratory and descriptive nature, involving the identification and integration of several open databases on population data, road routes, geolocation of the UBS in the municipality, and use of a geographic information system. As a result, we highlight the identification of areas of the municipality of Alegre with potential problems in accessing health care, achieved through an interactive web application implemented in Python that presents a heat map, with a granularity of 400m, according to the calculated health access index. The application can assist in health planning and management, identifying critical areas for better resource allocation and reducing inequalities. It can also support future planning, preventing problems in fast-growing regions and strengthening transparency in public management. As a continuation of the research, the integration of care history and ICDs would allow for more precise mapping of disease incidence, helping to combat diseases such as arboviruses.

Keywords: geospatial mapping; geographic information systems; population density; primary health care unit; public health.

1 INTRODUÇÃO

A transformação de dados em valor real, considerada um dos moes da Ciência de Dados, segundo Aalst (2016), é um desafio das instituições alinhadas ao fenômeno contemporâneo denominado de *data driven* (orientado a dados). Aliado a esse movimento, percebe-se que apenas o investimento em tecnologia, por parte das instituições, apesar de necessário, não é suficiente, pois há carência de profissionais com competências associadas à manipulação, investigação e interpretação de dados e que consigam aproveitar de forma plena as potencialidades das tecnologias inerentes a esse processo (Cristovão *et al.*, 2023). Essa carência pode ser amenizada com o desenvolvimento da competência em lidar com dados, denominada de *data literacy*, que, segundo Wolff *et al.* (2016), é a capacidade de fazer e responder perguntas associadas a problemas do mundo real a partir de bases de dados e por meio de processos de investigação.

A lacuna identificada em *data literacy* também engloba a área da saúde que, apesar de inúmeras iniciativas de inovação no uso de dados, demanda profissionais com competências para desenvolver soluções que efetivamente possam potencializar e melhorar a qualidade de atendimento e oferta de serviços à população. De acordo com o Ministério da Saúde (Brasil, 2024), a saúde envolve não apenas a cura e o tratamento de doenças, mas também a prevenção, a promoção de condições de vida saudáveis e a garantia do acesso a cuidados médicos de qualidade para todos os cidadãos. Dessa forma, o acesso à saúde, nos municípios, é organizado para atender as necessidades locais da população, sendo uma das principais estratégias de descentralização do Sistema Único de Saúde (SUS). O SUS estabelece que os municípios tenham um papel central na prestação de serviços de saúde, especialmente na atenção primária, pois é a esfera de governo mais próxima da população e melhor posicionada para responder às especificidades regionais.

No contexto municipal, o acesso à saúde é oferecido principalmente por meio das Unidades Básicas de Saúde (UBS) e das Equipes de Saúde da Família (ESF). Esse modelo de descentralização permite que os municípios tenham mais autonomia na execução das políticas de saúde, adaptando-as às demandas locais. Contudo, o acesso à saúde nos municípios pode ser impactado por fatores como o financiamento limitado, a carência de profissionais qualificados em áreas remotas, e as diferenças na infraestrutura entre regiões.

No contexto dos fenômenos *data driven* e *data literacy*, o projeto de extensão "Escritórios de Dados Municipais: difusão da cultura de dados para Cidades Inteligentes", da Universidade Federal do Espírito Santo (UFES), ao qual a presente pesquisa foi desenvolvida, parte da premissa de que as gestões municipais têm enfrentado rotineiramente enormes desafios para oferecer um conjunto de serviços públicos de qualidade aos munícipes, baseado em dados, em um cenário de recursos escassos e demandas crescentes da população.

O município de Alegre, localizado no sul do Estado do Espírito Santo, a cerca de 204 km da capital Vitória. Pertence à microrregião do Caparaó, com relevo montanhoso e clima tropical, favorecendo atividades agrícolas, como o cultivo de café e a pecuária. A população é estimada em cerca de 30 mil habitantes, com uma grande parcela em áreas rurais (Wikipedia, 2023). A sua administração é responsável pela gestão de serviços básicos, incluindo as Unidades Básicas de Saúde (UBS), que são distribuídas para atender tanto as zonas urbanas

quanto rurais. A proximidade com a Serra do Caparaó e a distribuição dos serviços de saúde no território tornam o mapeamento geoespacial relevante para entender o acesso dos munícipes às UBS.

Nesse contexto, observa-se que o acesso dos munícipes às UBS é um fator determinante para a utilização dos serviços de saúde e o consequente bem-estar da população. No entanto, muitos munícipes enfrentam barreiras que dificultam o acesso a essas unidades, especialmente quando consideradas a complexidade das rotas viárias.

Alinhado à problemática apresentada, o objetivo geral é identificar áreas do município de Alegre/ES que podem estar com problemas no acesso à saúde em função da distribuição das UBS em todo o território municipal, considerando-se a densidade populacional e a distância roteada das vias para acesso do munícipe à UBS mais próxima.

Acredita-se que, uma vez identificados pontos no município com maiores distâncias entre os munícipes e as UBS de forma proporcional à densidade populacional, seja possível auxiliar a tomada de decisão pelos gestores do município quanto a diversas ações relacionadas à facilidade de acesso aos cidadãos, tais como: (i) destinação de recursos para a organização e planejamento da construção de novas UBS; (ii) organizar e direcionar os atendimentos familiares nas residências conforme sinalização de locais mais densos no mapa, e (iii) verificação e melhoria das vias de acesso nos locais sinalizados pelo índice de acesso no mapa.

2 PROCEDIMENTOS METODOLÓGICOS

A pesquisa é de natureza aplicada, de abordagem quali-quantitativa, e de caráter exploratório e descritivo, envolvendo a identificação e integração de diversas bases de dados abertas sobre dados populacionais, vias rodoviárias e a geolocalização das 14 UBS do município de Alegre/ES, e o desenvolvimento de uma aplicação web interativa. Todos os dados primários utilizados na pesquisa são citados na próxima Subseção. A aplicação desenvolvida, encontra-se disponível no repositório de dados da pesquisa⁶.

⁶ Repositório de dados da pesquisa. Disponível em: https://github.com/labcidades-ufes/alegre_solucao_1.

2.1 Bases de Dados

Em busca de uma base de dados aberta, que apresentasse a população do município com uma granularidade fina, foi escolhido o portal *Humanitarian Data Exchange* (HDX)⁷ com a base denominada de “*Kontur Population: Global Population Density for 400m H3 Hexagons*”⁸. Essa base de dados representa os dados populacionais do mundo de forma espacial em uma grade de hexágonos com resolução de 400 metros. Essa tecnologia permite mapear a densidade populacional global para análises em escalas locais e globais, sendo essencial para garantir representações consistentes e precisas em áreas onde os dados populacionais podem ser escassos ou inexistentes (Kontur, 2023). Para áreas sem dados diretos, a *Kontur* combina fontes oriundas de censos, imagens de satélite e modelos algorítmicos que utilizam redes viárias e uso da terra. Métodos de interpolação distribuem populações de forma ponderada, garantindo estimativas consistentes mesmo em locais remotos, onde a combinação de fontes e metodologias torna a base de dados uma ferramenta poderosa para análises de densidade populacional (Kontur, 2023). Especialmente no contexto desta pesquisa, pelo fato da inexistência de dados oficiais da densidade populacional com granularidade adequada, a base da *Kontur* foi essencial para coletar a distribuição da população do município de Alegre.

O Sistema Integrado de Bases Geoespaciais do Estado do Espírito Santo (GEOBASES)⁹, que tem foco na padronização dos dados espaciais facilitando a análise geográfica e a integração de informações de diferentes fontes (Espírito Santo, c2025a), foi usado para acessar dados geográficos do município de Alegre como limites administrativos, rede viária e

⁷ *Humanitarian Data Exchange* (HDX) é um portal que facilita o acesso a dados humanitários, permitindo que organizações e indivíduos acessem, compartilhem e utilizem dados relacionados a crises humanitárias e desastres para melhor tomada de decisões. Ele oferece dados em diversos formatos, incluindo séries temporais, conjuntos de dados de emergência, e estatísticas demográficas. Disponível em: <https://data.humdata.org/>.

⁸ “*Kontur Population, Global Population Density for 400m H3 Hexagons*” é um serviço que foi originalmente criado pela empresa *Uber* (<https://www.uber.com/>). Contudo ele opera como uma tecnologia independente e de código aberto, permitindo que a *Kontur* aplique seu algoritmo globalmente, independentemente da cobertura de serviços da *Uber* (Kontur, 2023).

⁹ GEOBASES é um sistema que integra diversas bases de dados geoespaciais do ES atuando como uma ferramenta de transparência ativa e uma plataforma que reúne dados geográficos georreferenciados para diversos fins, como planejamento urbano, estudos ambientais e gestão de recursos (Espírito Santo, 2024a). Disponível em: <https://geobases.es.gov.br>.

localização de equipamentos públicos. Essas informações foram fundamentais para mapear a relação entre a distribuição territorial das UBS e a população do município.

O Instituto Jones dos Santos Neves¹⁰ (IJSN) disponibiliza uma ampla base de dados socioeconômicos e geoespaciais, sendo amplamente utilizado em projetos que exigem análises integradas e robustas tais como dados georreferenciados para análises espaciais detalhadas (Espírito Santo, c2025b). Para a presente pesquisa os dados extraídos no portal do IJSN foram fundamentais para o mapeamento geoespacial do acesso dos munícipes às UBS no município de Alegre, pois eles oferecem informações detalhadas e atualizadas sobre o território do município.

O *OpenStreetMap*¹¹ (OSM) é um projeto colaborativo global que oferece dados geográficos de acesso livre e gratuito, permitindo a criação, edição e manutenção de mapas detalhados, sendo amplamente utilizado para diversas finalidades, como navegação, planejamento urbano, análise geoespacial e resposta a emergências, devido à sua flexibilidade e acessibilidade (OpenStreetMap, 2025). Os dados da OSM foram fundamentais para o mapeamento geoespacial das vias para a realização do roteamento dos munícipes às UBS no município de Alegre.

2.2 Índice de Acesso à Saúde

Para viabilizar a exploração dos dados geoespaciais coletados, bem como realizar experimentações visuais sobre o mapa de calor que iria apresentar o acesso às UBS em função da densidade populacional, foi utilizado um Sistema de Informação Geográfica¹² (SIG) de

¹⁰ Instituto Jones dos Santos Neves é uma instituição de pesquisa vinculada à Secretaria de Estado de Economia e Planejamento do Estado do Espírito Santo. Seu objetivo principal é fornecer informações que apoiem o planejamento e a execução de políticas públicas, abrangendo aspectos como demografia, economia, meio ambiente e infraestrutura (Espírito Santo, 2024b).

¹¹ OpenStreetMap é uma plataforma mantida pela OpenStreetMap Foundation, uma organização sem fins lucrativos que apoia o desenvolvimento e a infraestrutura do projeto, garantindo a disponibilidade contínua e a atualização de seus dados. Disponível em: <https://www.openstreetmap.org/>.

¹² Sistema de Informação Geográfica (SIG) ou *Geographic Information System* (GIS) é um sistema de hardware, software, informação espacial, procedimentos computacionais e recursos humanos que permite e facilita a análise, gestão ou representação de informação geográfica (Wikipedia, 2024).

código aberto denominado QGIS¹³, principalmente por permitir tais experimentações de forma interativa e exploratória.

As experimentações no QGIS foram realizadas em busca de um índice que pudesse ser adequadamente traduzido para um formato visual em um mapa de calor com granularidade dos hexágonos lidos na base Kontur (2023). O cálculo do índice foi aplicado para cada hexágono e tomou como base duas variáveis: a densidade populacional do hexágono e a distância roteada do centro do hexágono (centroide) até a UBS mais próxima.

Para a adequada utilização das duas variáveis no índice, sabendo-se que estas representam grandezas distintas, com unidades e escala próprias, foram utilizadas as variáveis padronizadas (subtraídas de suas médias e divididas pelo desvio padrão), obtendo-se, para ambas as medidas, valores com média '0' e desvio padrão '1', conforme mostra a Equação 1. Esta transformação evita que uma das variáveis ocasionalmente domine a métrica do índice.

$$Z = \frac{X - \bar{X}}{s} \quad (1)$$

Onde:

X variável original;

$\bar{X}$ média da variável;

s desvio padrão da variável

Z variável transformada, com média = '0' e desvio padrão = '1'.

Contudo, o uso da variável transformada Z implica em observações negativas. Portanto, para adequada construção do índice, os resultados foram deslocados da média em 10 unidades, tornando todas as observações maiores que 0 e permitindo uma interação com propriedades matemáticas adequadas entre as variáveis.

Seja Z a variável aleatória, cuja esperança é dada por $E(Z) = \bar{X}$ e a variância $Var(Z) = s^2$. Seja ainda c uma constante qualquer, conforme as Equações 2 e 3.

$$E[Z + c] = E[Z] + E[c] = c + E[Z] = c + \bar{X} \quad (2)$$

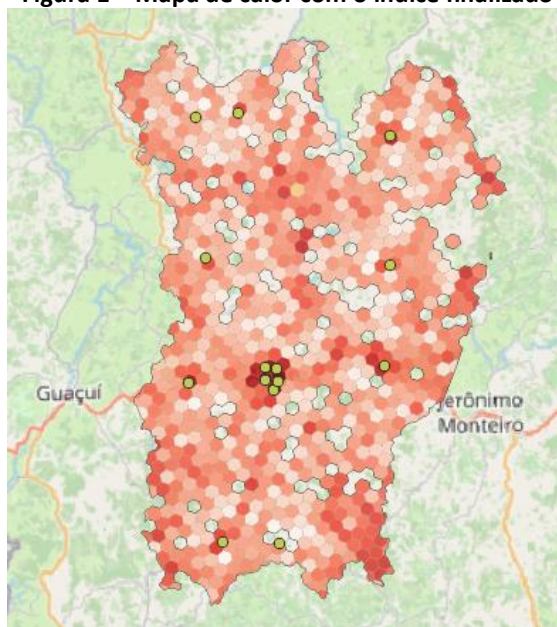
¹³ Quantum GIS (QGIS) é uma ferramenta *open source* SIG compatível com Windows, Mac, Linux, mantida por uma comunidade global de desenvolvedores, colaboradores e organizações. Disponível em: <https://qgis.org/>.

$$Var[Z + c] = Var[Z] = s^2 \quad (3)$$

Logo, o deslocamento da variável observada não afeta o desvio da mesma, garantindo que a dispersão entre as observações seja mantida. Além disso, as variáveis podem ser combinadas (somadas ou multiplicadas) de forma justa e balanceada, uma vez que a contribuição de cada uma no índice é proporcional à sua variação relativa.

O resultado do experimento pode ser visto no mapa da Figura 1. Nesse mapa, as unidades de saúde são representadas por pontos na cor bege e os hexágonos possuem variação de cor proporcional ao índice calculado.

Figura 1 – Mapa de calor com o índice finalizado



Fonte: Elaboração própria, recorte feito a partir da interface do software QGIS.

Apesar do valor de 10 ter sido uma decisão arbitrária, não afetou a variância do índice, uma vez que o desvio padrão independe da média. Ou seja, embora a média tenha sido arbitrariamente deslocada, a variância não foi afetada e, portanto, não houve prejuízo para a interpretabilidade do índice. Por fim, calibrou-se o modelo para que o índice fosse igual ao produto entre as duas variáveis, isto é, o índice igual ao produto da distância padronizada deslocada pela população padronizada deslocada.

2.3 Desenvolvimento da Aplicação

O mapa de calor com o índice de acesso à saúde obtido no *software QGIS* precisou ser implementado na linguagem de programação *Python* para que a aplicação ficasse independente para ser usada pelos gestores da prefeitura pelo simples acesso a um navegador web. Além disso, aproveitou-se para criar interatividade na escolha da exibição de camadas no mapa em busca de facilitar a compreensão do processo de obtenção do resultado pelo gestor: limites do município, sistema viário, população, localização dos centróides, peso dos hexágonos e rotas calculadas. O Quadro 1 mostra as principais bibliotecas *Python* usadas no desenvolvimento da aplicação.

Quadro 1 – Bibliotecas Python usadas na implementação

Download de dados:
• requests: para fazer as requisições - https://pypi.org/project/requests/.
Tratamento e processamento dos dados:
• geopandas: para manipulação de geodataframes - https://geopandas.org. • osmnx: para criar o grafo viário com os nós e calcular os nós mais próximos aos centróides e as unidades de saúde- https://osmnx.readthedocs.io. • networkx: para calcular a menor rota entre os centróides e as unidades de saúde - https://networkx.org.
Exibição dos dados como app do shiny:
• folium: para criação de mapas interativos - https://pypi.org/project/folium. • shiny: para criação de web app interativos - https://shiny.posit.co/py.

Fonte: Elaboração própria.

Todos os dados utilizados na pesquisa, inclusive o código fonte da aplicação desenvolvida, encontra-se disponível no repositório de dados da pesquisa¹⁴. O acesso à execução da aplicação publicada na web encontra-se em link do LabCidades¹⁵.

3 RESULTADOS E DISCUSSÃO

A identificação das áreas do município de Alegre com potenciais problemas no acesso à saúde foi alcançada por intermédio da aplicação desenvolvida que apresenta um mapa de

¹⁴ Repositório de dados da pesquisa. Disponível em: <https://github.com/labciudades-ufes/>.

¹⁵ Aplicação web desenvolvida. Disponível para execução em: http://labciudades.ufes.br/alegre_saude/.

calor em função do índice de acesso à saúde calculado, conforme mostra a Figura 2. Nessa figura observa-se a interface inicial da aplicação com interação para escolha das camadas desejadas. Nesse caso, estão selecionadas as camadas dos limites da cidade de Alegre e o peso dos hexágonos coloridos em função do índice calculado. Sendo que a coloração avermelhada mais escura dos hexágonos representa um pior acesso do munícipe ao posto de saúde, e a mais clara representa um acesso melhor. Em todos os casos, para essa coloração, considerou-se as variáveis da densidade populacional e a distância roteada pelas vias rodoviárias até a unidade de saúde mais próxima.

Associado a todo o processo de desenvolvimento desse resultado houve também um significativo amadurecimento de competências do tipo *data literacy* pelos participantes do projeto ao exercerem um papel investigativo na descoberta de bases de dados abertas úteis para a solução e no lido com diversas tecnologias e suas peculiaridades, em especial aquelas inerentes a processos de geoespacialização. Considera-se também, enquanto desenvolvimento de *data literacy* pelos gestores da prefeitura, a desmistificação do uso de tecnologias e o uso de bases de dados abertas na solução de diversos problemas em outros contextos.

Segundo Hjørland (2018), os dados nunca são independentes de contexto, mas sempre produzidos a partir de determinadas perspectivas, que contêm critérios próprios de relevância. Assim, é importante considerar que a informação sobre a facilidade de acesso à saúde, revelada em cada hexágono no mapa da aplicação, representa um conjunto de dados que deve ser devidamente interpretado pelos gestores da área da saúde do município, pois eles são os profissionais que compreendem o contexto de forma holística e podem considerar outras variáveis relevantes no processo de análise e tomada de decisões.

Ainda sim é possível inferir alguns usos com potencial de impacto positivo no município, tanto no planejamento estratégico quanto na gestão de recursos destinados à saúde, como por exemplo a identificação de áreas prioritárias. Pois, com o mapa de calor, é possível identificar áreas críticas, onde o acesso aos serviços é mais difícil e/ou são áreas com alta demanda. Tal leitura facilita na tomada de decisões para alocação de recursos e priorização de intervenções, como melhorias na malha viária, transporte público, etc. Ao

identificar essas barreiras de acesso, o município pode mitigar disparidades no acesso aos serviços de saúde, promovendo maior equidade nos cuidados aos cidadãos.

Figura 2 – Interface da aplicação desenvolvida



Fonte: Elaboração própria.

Também é possível utilizar a aplicação desenvolvida como um apoio ao planejamento de médio e longo prazo, pois a visualização do índice também fornece subsídios para, por exemplo, orientar ações preventivas em regiões com crescimento populacional acelerado ou com risco de emergências de saúde. Além disso, a ferramenta pode fortalecer a transparência e a tomada de decisão baseada em evidências por parte da gestão do município, melhorando o relacionamento com a sociedade civil, à medida que reforça a confiança na administração pública.

4 CONSIDERAÇÕES FINAIS

O objetivo geral foi atingido, uma vez que houve a identificação e mapeamento das áreas do município de Alegre/ES que podem estar com problemas no acesso à saúde em função da distribuição das UBS na área total no município, considerando-se a densidade populacional e a distância roteada das vias para acesso do munícipe à UBS mais próxima. A solução levou em consideração as variáveis: densidade populacional com granularidade de hexágonos de 400 m e a distância roteada pelas vias rodoviárias até a unidade de saúde mais próxima.

Os participantes do projeto ampliaram suas competências associada a *data literacy* ao investigar bases de dados abertas e lidar com tecnologias, especialmente as de geoespacialização. Enquanto os gestores da prefeitura puderam desmistificar o uso de tecnologias e reconhecer o potencial dessas bases para resolver problemas em diversos contextos.

Apesar da aplicação ainda não ter sido efetivamente usada pela prefeitura, é possível enxergar aplicabilidade no planejamento estratégico e na gestão de recursos da saúde, auxiliando na identificação de áreas críticas com difícil acesso aos serviços ou alta demanda. Isso facilita a alocação de recursos e a priorização de intervenções, reduzindo desigualdades no atendimento. Além disso, a aplicação pode apoiar o planejamento de médio e longo prazo, ajudando a prevenir problemas em regiões com crescimento populacional acelerado ou risco de emergências. Também fortalece a transparência e a tomada de decisões baseadas em evidências, melhorando a relação entre a gestão municipal e a sociedade.

Como continuidade do trabalho e considerando o acesso ao histórico de atendimentos de saúde dos munícipes e seus respectivos CIDs¹⁶ integrados na aplicação, seria possível fazer um mapeamento mais preciso sobre a incidência de determinadas doenças de forma geolocalizada. Como, por exemplo, o combate a doenças infectocontagiosas, ou o combate a doenças transmitidas por mosquitos, como as arboviroses dengue, chikungunya e zika.

REFERÊNCIAS

AALST, Wil Van der. Data science in action. In: AALST, Wil Van der (org.). **Process mining**. Berlin: Springer, 2016. p. 3-23. DOI: [10.1007/978-3-662-49851-4_1](https://doi.org/10.1007/978-3-662-49851-4_1). Disponível em: https://doi.org/10.1007/978-3-662-49851-4_1. Acesso em: 28 fev. 2025.

BRASIL. Ministério da Saúde. SUS. **Sistema Único de Saúde (SUS)**. 2024. Disponível em: <https://www.gov.br/saude/pt-br/sus>. Acesso em: 28 fev. 2025.

CRISTOVÃO, Henrique Monteiro et al. Práticas em análise de redes sociais para Data Literacy em Ciência de Dados no âmbito da formação na área Ciência da Informação. In: ENCONTRO NACIONAL DE PESQUISA E PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO, 23., 2023, Aracaju. **Anais [...]**. Aracaju: ANCIB, 2023. Disponível em:

¹⁶ CID é a sigla para Classificação Internacional de Doenças, uma ferramenta da Organização Mundial de Saúde (OMS).

<https://enancib.ancib.org/index.php/enancib/xxxiiienancib/paper/view/1778>. Acesso em: 28 fev. 2025.

ESPÍRITO SANTO (Estado). Secretaria de Estado de Economia e Planejamento. **Instituto Jones dos Santos Neves**. c2025b. Disponível em: <https://ijsn.es.gov.br/>. Acesso em: 28 fev. 2025.

ESPÍRITO SANTO (Estado). Secretaria de Gestão e Recursos Humanos. **Sistema integrado de Bases Geoespaciais do Estado do Espírito Santo (Geobases)**. c2025a. Disponível em: <https://geobases.es.gov.br/>. Acesso em: 28 fev. 2025.

HJØRLAND, Birger. Data (with big data and database semantics). **Knowledge Organization**, [S. l.], v. 45, n. 8, p. 685-708, 2018. Disponível em: <https://www.isko.org/cyclo/data>. Acesso em: 28 fev. 2025.

KONTUR. **Humanitarian Data Exchange (HDX)**. [Data set]. 2023. Disponível em: <https://data.humdata.org/dataset/kontur-population-dataset>. Acesso em: 28 fev. 2025.

OPENSTREETMAP. **OpenStreetMap**. 2025. Disponível em: <https://www.openstreetmap.org/>. Acesso em: 28 fev. 2025.

WIKIPEDIA. **Alegre (Espírito Santo)**. 2023. Disponível em: [https://pt.wikipedia.org/wiki/Alegre_\(Esp%C3%ADrito_Santo\)](https://pt.wikipedia.org/wiki/Alegre_(Esp%C3%ADrito_Santo)). Acesso em: 28 fev. 2025.

WIKIPEDIA. **Sistema de informação geográfica**. 2024. Disponível em: https://pt.wikipedia.org/wiki/Sistema_de_informa%C3%A7%C3%A3o_geogr%C3%A1fica. Acesso em: 28 fev. 2025.



[4.0 Internacional](https://creativecommons.org/licenses/by/4.0/)

Esta licença permite compartilhamento, remixe, adaptação e criação a partir do trabalho, mesmo para fins comerciais, desde que sejam atribuídos créditos ao(s) autor(es). Conteúdos elaborados por terceiros, citados e referenciados nesta obra não são cobertos pela licença.

Como citar este trabalho:

CRISTOVÃO, Henrique Monteiro; MOURA, Alexsander Barbosa; SILVA, Renan Carlos da; CRISTOVÃO, Ágata; ABREU, Camile Pereira de. Mapeamento geoespacial de áreas com potenciais problemas no acesso às unidades básicas de saúde no município de Alegre/ES. In: WORKSHOP DE INFORMAÇÃO DADOS E TECNOLOGIA, 8., 2025, Marília, SP. **Anais [...]**. Marília, SP: Universidade de Marília, 2025. DOI: <https://doi.org/10.22477/viii.widat.276>.



WIDaT 2025

VIII WORKSHOP DE INFORMAÇÃO,
DADOS E TECNOLOGIA

MARÍLIA • SP

Utilizando o ChatGPT como ferramenta de suporte para a construção de uma base de dados textual sobre violências: apresentação e análise do processo de desenvolvimento

Using ChatGPT as a support tool for building a textual database on violence: presentation and analysis of the development

Guilherme Ataíde Dias¹



<http://lattes.cnpq.br/9553707435669429>



<https://orcid.org/0000-0001-6576-0017>

Tassyara Onofre de Oliveira²



<http://lattes.cnpq.br/5328618364332415>



<https://orcid.org/0000-0003-4048-8322>

Arthur Antônio Martins Sales Campelo³



<http://lattes.cnpq.br/2370276793833303>



<https://orcid.org/0000-0001-5907-917X>

Vivian Roberta Lima de Carvalho⁴



<http://lattes.cnpq.br/4602577185970003>



<https://orcid.org/0009-0003-6645-7660>

Resumo

Este estudo descreve a construção de uma base de dados textual de referência sobre violências, com o objetivo de aprimorar a recuperação de informações para pesquisadores da Rede Interdisciplinar de Estudos sobre Violências. A pesquisa, de caráter aplicado e abordagem qualitativa, integrou princípios da Ciência da Informação e tecnologias emergentes, utilizando o ChatGPT como ferramenta de suporte na criação de uma página web, tesouro temático, gerador de itens e imagens ilustrativas. O repositório foi

¹ Doutor em Comunicação (Ciência da Informação), Universidade de São Paulo (USP), São Paulo, Brasil. Professor Titular, Universidade Federal da Paraíba, João Pessoa, Brasil. guilhermeataide@ccsa.ufpb.br.

² Doutora em Ciência da Informação, Universidade Federal da Paraíba (UFPB), João Pessoa, Paraíba, Brasil. Pós-doutoranda júnior bolsista do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Programa de Pós-Graduação em Ciência da Informação da Universidade Federal da Paraíba (UFPB), João Pessoa, Paraíba, Brasil. Estância de pesquisa na Universitat de València (UV), València, Comunidade Valenciana, Espanha. tassyaraonofre@gmail.com.

³ Graduando em Biblioteconomia, Universidade Federal da Paraíba (UFPB), João Pessoa, Brasil. Bolsista de Iniciação Científica do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), João Pessoa, Paraíba, Brasil. arthur.campelo@academico.ufpb.br.

⁴ Graduanda em Biblioteconomia, Universidade Federal da Paraíba (UFPB), João Pessoa, Paraíba, Brasil. Bolsista de Iniciação Científica do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), João Pessoa, Paraíba, Brasil. viivacarvalho@gmail.com.

implementado no DSpace 7, organizado em coleções temáticas aderentes aos princípios FAIR, resultando em maior eficiência na recuperação de dados. Os resultados indicam que a inteligência artificial reduziu significativamente o tempo de desenvolvimento, embora ajustes humanos sejam essenciais para garantir qualidade. Conclui-se que o uso do ChatGPT viabilizou avanços no projeto, mas a automação de coleta e ingestão de metadados e a capacitação em DSpace permanecem como desafios a serem superados.

Palavras-chave: Inteligência Artificial; ChatGPT; DSpace; base de dados; violência.

Abstract

This study describes the development of a reference textual database on violence aimed at enhancing information retrieval for researchers from the Interdisciplinary Network for Violence Studies. Employing an applied research approach with a qualitative focus, it integrated Information Science principles and emerging technologies, using ChatGPT as a support tool to create a web page, thematic thesaurus, item generator, and illustrative images. The repository, implemented in DSpace 7 and organized into FAIR-compliant thematic collections, improved data retrieval efficiency. Results demonstrate that artificial intelligence significantly reduced development time, though human adjustments remain necessary to ensure quality. The findings suggest that ChatGPT facilitated project advancements, yet challenges such as automating metadata collection and ingestion, along with expertise in DSpace customization, persist. This work highlights the potential of artificial intelligence in database development while underscoring the need for ongoing refinement.

Keywords: Artificial Intelligence; ChatGPT; DSpace; database; violence.

1 INTRODUÇÃO

O desenvolvimento da inteligência artificial ocorreu de forma paulatina ao longo das últimas décadas, tendo como ponto de partida o período iniciado na Segunda Guerra Mundial. Esse avanço alternou períodos de intensa atividade de pesquisa com fases de menor produção científica, influenciados tanto pelos resultados obtidos pelos pesquisadores quanto pelos diversos interesses das indústrias e dos governos financiadores.

De acordo com Russel e Norvig (2004), o primeiro trabalho atualmente reconhecido como IA foi desenvolvido por Warren McCulloch e Walter Pitts em 1943. Nesse estudo, os autores propuseram um modelo matemático de neurônios artificiais baseado na lógica booleana. Esse trabalho foi fundamental para a criação das redes neurais artificiais e serviu como base para o desenvolvimento da IA, especialmente no campo do aprendizado de máquina. Outro pioneiro da área foi o cientista britânico Alan Turing, que, em 1950 publicou um artigo seminal intitulado de “*Computing Machinery and Intelligence*” no qual explorou a possibilidade de máquinas exibirem comportamento inteligente (Russel; Norvig, 2004). O verdadeiro marco da IA como disciplina científica formal ocorreu em 1956, na Conferência de *Dartmouth*, organizada por John McCarthy,

Marvin Minsky, Nathaniel Rochester e Claude Shannon. Nesse evento, o termo “Inteligência Artificial” foi cunhado (Russel; Norvig, 2004).

Atualmente, constata-se um grande avanço no desenvolvimento e uso da IA, impulsionado principalmente pelas redes neurais profundas e pela ampla disponibilidade de dados. A crescente proliferação de modelos de linguagem de grande porte (LLMs), como *ChatGPT* e outras ferramentas no mercado, tem alavancado tanto a pesquisa acadêmica quanto as iniciativas comerciais. Em paralelo, melhorias na infraestrutura computacional e na capacidade de processamento viabilizam treinamentos cada vez mais robustos e especializados. Entretanto, a disseminação veloz dessas tecnologias evidencia a necessidade de regulamentações adequadas, bem como de reflexões éticas e práticas de governança que assegurem sua utilização responsável.

A investigação apresentada neste estudo descreve o processo de desenvolvimento de uma base de dados textual de referência sobre a temática das “violências”, com o objetivo de apoiar pesquisadores que atuam nesse campo, especialmente aqueles vinculados à *Rede Interdisciplinar de Estudos sobre Violências* (Universidade Federal da Paraíba, 2025).

A motivação para o desenvolvimento da base de dados textual de referência sobre a temática das “violências” parte do pressuposto de que a pesquisa realizada em uma base de dados temática, cujos documentos passaram previamente por um processo criterioso de seleção dentro do domínio de interesse, tende a apresentar maior qualidade nos resultados recuperados em comparação à pesquisa realizada em bases que abrangem temáticas diversas.

Por se tratar de uma base de dados textual de referência, optamos por não armazenar os conteúdos originais na sua íntegra - sejam eles artigos, dissertações, teses, notícias ou outros tipos de documento. Essa decisão foi motivada principalmente pela necessidade de evitar conflitos relacionados aos direitos autorais. Dessa forma, a base de dados reúne apenas os elementos descritores e as referências das obras, materializados em dois arquivos distintos: um no formato texto simples e outro estruturado na linguagem XML.

Um ponto fundamental considerado no desenvolvimento da base de dados foi torná-la o mais aderente possível aos princípios FAIR (*Findable, Accessible, Interoperable* e *Reusable*), com o objetivo de ampliar o uso, o reuso e o compartilhamento dos dados disponibilizados (Wilkinson *et al.*, 2016). A adoção desse padrão garante maior visibilidade e facilidade na recuperação dos recursos, além de contribuir para o fortalecimento da colaboração científica. Ao seguir essas diretrizes, a base de dados proporciona aos pesquisadores condições mais adequadas para o acesso, a interpretação e a reutilização das informações, potencializando, assim, o impacto das pesquisas relacionadas à temática das “violências”.

2 PROCEDIMENTOS METODOLÓGICOS

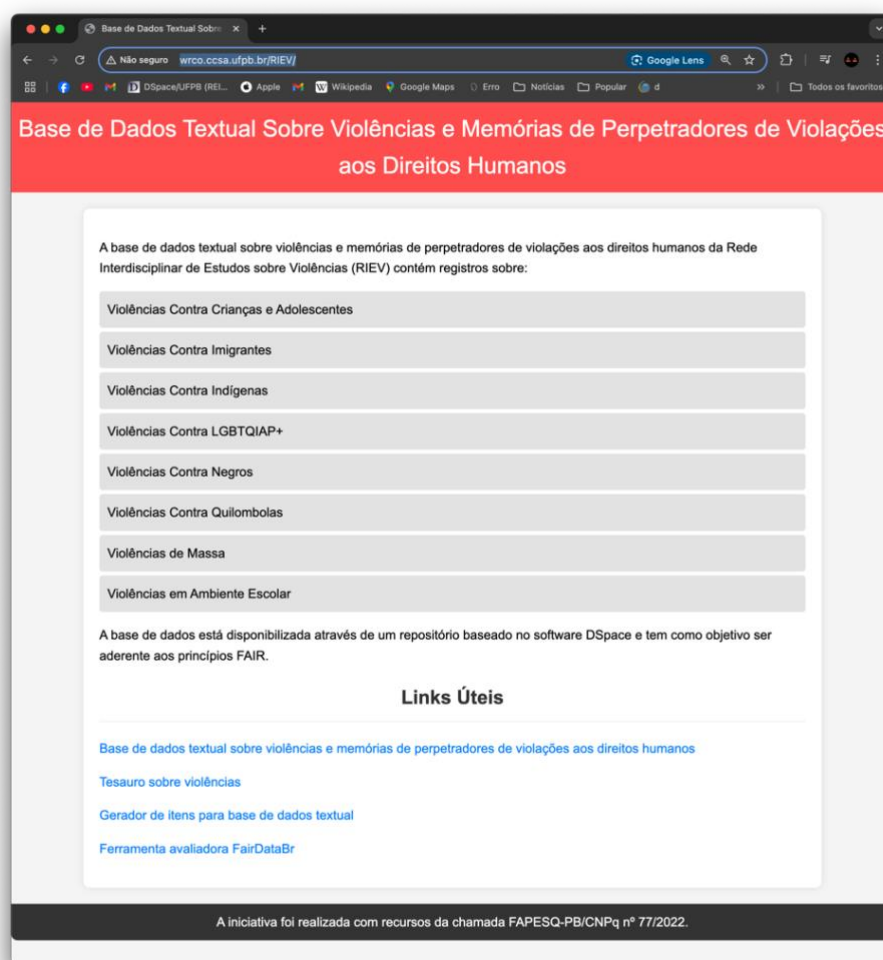
A metodologia adotada para o desenvolvimento da base de dados textual de referência sobre violências caracteriza-se como pesquisa aplicada, pois tem como objetivo a criação de um sistema para apoiar pesquisadores na recuperação e organização de informações sobre o tema. A abordagem utilizada é predominantemente qualitativa, centrada na estruturação e depósito dos dados. O trabalho integra tecnologias emergentes, automação de processos e princípios da Ciência da Informação, proporcionando maior eficiência na gestão e disseminação do conhecimento sobre violências.

O *ChatGPT* foi utilizado como uma ferramenta de suporte para a criação de instrumentos necessários à operacionalização da base de dados. A seguir, indicamos quais foram: 1) Página na *Web* para acesso a base de dados e demais ferramentas de suporte; 2) Tesouro sobre violências; 3) Gerador de itens para a base de dados e 4) Imagens ilustrativas das coleções e itens da base de dados. Utilizamos a versão com assinatura do *ChatGPT*, que emprega o modelo *GPT-4*.

A página *Web* para acesso à base de dados e demais ferramentas de suporte foi criada com auxílio do *ChatGPT*, por meio de um *prompt* contendo instruções textuais detalhadas sobre como gostaríamos que ela fosse estruturada e apresentada. O resultado foi obtido na forma de um arquivo contendo código HTML, posteriormente testado localmente em um navegador. Durante o desenvolvimento foram realizadas diversas

interações com o *ChatGPT* até que chegássemos a uma versão plenamente às nossas expectativas iniciais. Essa abordagem mostrou-se bastante eficiente, uma vez que possibilitou ajustes rápidos e pontuais, garantindo, assim, maior controle sobre o resultado final. A Figura 1 ilustra o código gerado, exibido em um navegador *Web* sob a forma de página, destacando a clareza e a funcionalidade obtidas por meio dessa abordagem interativa.

Figura 1 – Página de acesso aos instrumentos necessários à operacionalização da base de dados



Fonte: Elaboração própria.

Para realizar o depósito dos registros sobre violência na base de dados, verificamos a necessidade de desenvolver um vocabulário controlado, especificamente um tesauro, para garantir um controle terminológico eficiente das palavras-chave associadas a cada item inserido. Essa estratégia contribui bastante para aumentar a

precisão na recuperação dos dados, reduzindo ambiguidades e facilitando o acesso pelos usuários.

O uso do tesauro permite estabelecer uma padronização consistente na descrição dos conteúdos, ampliando as possibilidades de integração com outras bases de dados e favorecendo futuras ações de compartilhamento e colaboração científica.

A extração de termos a partir de um tesauro é uma etapa fundamental para a indexação de documentos na base de dados. Com a organização criteriosa dos termos, consolidou-se um vocabulário controlado que facilita a recuperação da informação. Esse processo assegurou uma indexação inicial padronizada e reduziu significativamente o esforço humano na categorização dos conteúdos, otimizando o tempo de ingestão de dados na base.

Os termos iniciais para a criação do vocabulário controlado foram extraídos a partir de títulos e resumos associados de 105 documentos selecionados previamente com temas relacionados a diversos tipos de “violências”. Os títulos e resumos foram submetidos ao *ChatGPT* por meio de um arquivo de texto. Solicitamos através do *prompt* a geração de um arquivo em formato DOCX para *download* que reunisse em uma lista os termos mais representativos de “violências” contidos nos respectivos títulos e resumos submetidos. A lista de termos representativos das “violências” foi utilizada como um base norteadora inicial para a construção de um tesauro por pesquisadores na área da Ciência de Informação. Para a construção do tesauro foi realizada com auxílio da ferramenta *Thesa*⁵ (Tesauro Semântico Aplicado) (Gabriel Junior; Laipelt, 2017), desenvolvida por pesquisadores da Universidade Federal do Rio Grande do Sul.

O desenvolvimento de um formulário para a geração dos elementos descritivos dos recursos informacionais sobre “violências” (itens da base de dados) mostrou-se necessário, visto que o que é efetivamente depositado no *DSpace* não é o recurso informacional fonte propriamente dito (artigo, dissertação, teses, notícia ou outro tipo de documento), mas sim uma descrição desse recurso. Essa característica configura a base como uma de dados textual de referência.

⁵ Disponível em: <https://www.ufrgs.br/tesauros/index.php/thesa>. Acesso em: 19 maio 2025.

Assim como na construção da página para acesso à base de dados, nesse caso também utilizamos o *ChatGPT* em diversas interações até alcançar uma configuração adequada do formulário de entrada de dados. Cabe mencionar que alguns ajustes no código HTML do formulário precisaram ser realizados manualmente, até obtermos uma versão que atendesse plenamente aos nossos propósitos. O uso do formulário possibilitou gerar os itens a serem depositados na base de dados com maior agilidade e melhor controle de qualidade.

Após a captura dos dados pelo formulário, estes são exportados para dois arquivos antes de serem depositados no *DSpace*: um arquivo em formato de texto contendo os principais dados do recurso e um segundo, também em formato texto, porém estruturado na Linguagem XML. A Figura 2 exemplifica o conteúdo de um dos arquivos gerados com os descritores do recurso informacional, no caso a versão em XML.

Figura 2 – Item da base de dados (versão XML)

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:dc="http://purl.org/dc/elements/1.1/">
  <rdf:Description rdfabout="http://example.org/resource/1">
    <dc:title>Desconstruindo os arquivos da violência escolar: reflexões sobre a relação entre a docência e a Filosofia da Educação</dc:title>
    <dc:type>Artigo</dc:type>
    <dc:creators>Amarildo Luiz Trevisan</dc:creators>
    <dc:keywords>Violência simbólica</dc:keywords>
    <dc:keywords>Preconceito</dc:keywords>
    <dc:keywords>Exclusão social</dc:keywords>
    <dc:keywords>Violência psicológica</dc:keywords>
    <dc:coverage>Ponta Grossa</dc:coverage>
    <dc:date>20/07/2023</dc:date>
    <dc:format>application/pdf</dc:format>
    <dc:language>pt</dc:language>
    <dc:relation>n/a</dc:relation>
    <dc:source>Práxis Educativa</dc:source>
    <dc:doi>https://doi.org/10.5212/PraxEduc.v.18.22099.068</dc:doi>
    <dc:issn>1809-4031</dc:issn>
    <dc:orcid>https://orcid.org/0000-0002-3575-4369</dc:orcid>
    <dc:link>https://revistas.uepg.br/index.php/praxiseducativa/article/view/22099</dc:link>
    <dc:reference>TREVISAN, A. L. Desconstruindo os arquivos da violência escolar: reflexões sobre a relação entre a docência e a Filosofia da Educação. Práxis Educativa, [S. l.], v. 18, p. 1-19, 2023. DOI: 10.5212/PraxEduc.v.18.22099.068. Disponível em: https://revistas.uepg.br/index.php/praxiseducativa/article/view/22099.</dc:reference>
    <dc:description>Este texto apresenta uma reflexão crítica sobre os arquivos da violência escolar como um problema filosófico e psicanalítico. Utilizando a expressão "mal de arquivo", de Derrida, como ponto de partida, discutem-se os movimentos da docência diante do problema da violência escolar. A partir de uma experiência de pesquisa, são apresentadas as opiniões de professores sobre as dificuldades que enfrentam no contexto escolar em relação à violência. Com base na hermenêutica desconstrutiva, o texto compreende a escola como um espaço de enfrentamento de estereótipos, preconceitos e fantasmas sobre a violência, facilitando a convivência com as diferenças. Conclui-se pela necessidade de pensar o arquivo escolar como uma "tela desconstrucionista", capaz de romper com o individualismo e a indiferença em relação ao outro. Palavras-chave: Mal de arquivo. Filosofia da Educação. Violência escolar.</dc:description>
    <dc:abstract>This text presents a critical reflection on the archives of school violence as a philosophical and psychoanalytical problem. Using Derrida's expression "archive fever" as a starting point, it discusses teaching movements against the school violence issue. From a research experience, teachers' opinions about the difficulties they face in the school context in relation to violence are presented. Based on the deconstructive hermeneutics, the text understands the school as a space for confronting stereotypes, prejudices and ghosts about violence, facilitating coexistence with differences. It is concluded that there is a need to think of the school archive as a "deconstructionist canvas", capable of breaking with individualism and indifference towards the other.</dc:abstract>
    <dc:extent>19 p.</dc:extent>
    <dc:quality>n/a</dc:quality>
    <dc:rights>CC BY 4.0</dc:rights>
  </rdf:Description>
</rdf:RDF>
```

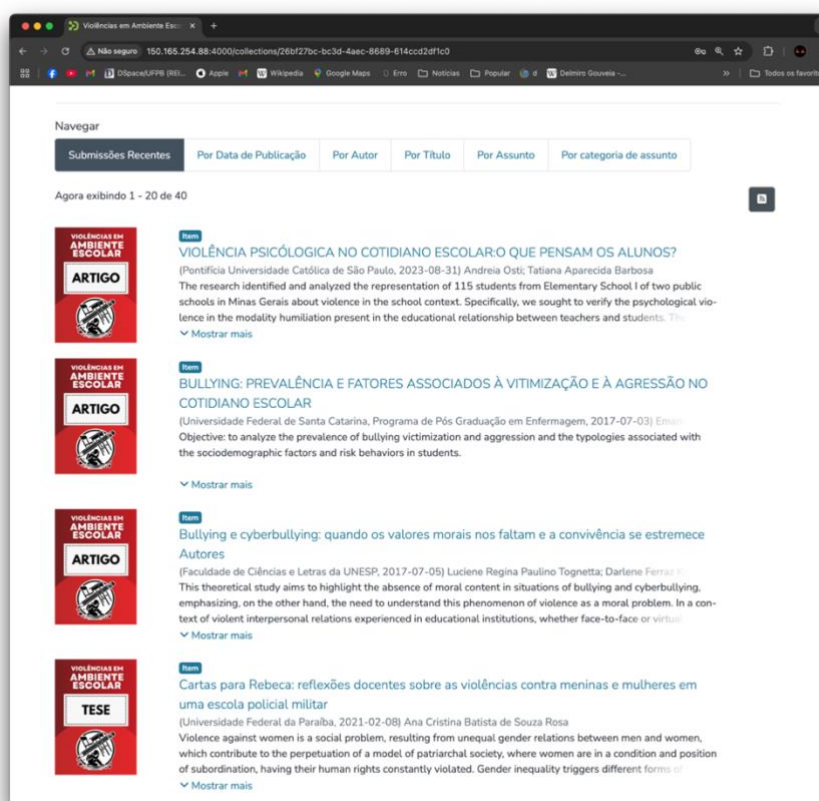
Fonte: Elaboração própria.

As imagens ilustrativas das coleções e itens da base de dados (*thumbnails*) também foram geradas através do *ChatGPT*. Cada coleção e cada item cadastrado na base possuem uma imagem específica que os representa visualmente, facilitando a identificação rápida pelos usuários. Essa abordagem tornou possível obter uma padronização estética e funcional nas representações gráficas, contribuindo para uma interface mais intuitiva.

O produto de *software* selecionado para armazenar e disponibilizar os registros da base de dados foi o *DSpace 7*. Essa versão foi escolhida porque, no momento da instalação, era a versão de produção recomendada (*current release*). Embora o *DSpace 8* já estivesse disponível, ainda não era considerada a opção mais estável. O *DSpace 7* foi instalado de forma containerizada em um servidor *Linux Ubuntu* virtualizado nas dependências da Superintendência de Tecnologia da Informação da Universidade Federal da Paraíba (STI/UFPB).

Para estruturar logicamente a base de dados, criamos uma comunidade *DSpace* intitulada *Violências e Memórias - Rede Interdisciplinar de Estudos sobre Violências (RIEV)*. Dentro desta comunidade criamos oito coleções, a saber: 1-*Violências Contra Crianças e Adolescentes*, 2-*Violências Contra Imigrantes*, 3-*Violências Contra Indígenas*, 4-*Violências Contra LGBTQIAPN+*, 5-*Violências Contra Negros*, 6-*Violências Contra Quilombolas*, 7-*Violências de Massa* e 8-*Violências em Ambiente Escolar* (Figura 3).

Figura 3 – Comunidade e Coleções da base de dados



Fonte: Elaboração própria.

Para ampliar a aderência dos dados disponibilizados ao padrão FAIR, utilizamos a ferramenta de avaliação de conjuntos de dados *FairDataBR* (*FairDataBR*, 2025) a fim de verificar seu grau de conformidade. Indicamos também que os elementos de metadados de cada item depositado na base de dados são compatíveis com o padrão *Dublin Core*.

3 RESULTADOS

A base de dados textual de referência sobre violências está operacional em fase de testes de povoamento, encontra-se disponível para consulta pela comunidade no endereço <http://150.165.254.88:4000/home> (Figura 4).

Figura 4 – Exemplos de itens depositados na base de dados



Fonte: Elaboração própria

Uma das possibilidades de aprimoramento no processo de construção da base de dados refere-se à demora na coleta de artigos, dissertações, teses, notícias e outros documentos para a extração de seus descritores pelos pesquisadores. Esse desafio poderia ser mitigado por meio da implantação de um *harvester* compatível com OAI-PMH,

permitindo a coleta automatizada de metadados diretamente dos portais de periódicos, o que é particularmente relevante, considerando que os artigos representam a maior parte dos recursos informacionais depositados na base.

De forma análoga, o processo de ingestão (*ingestion*) dos itens na base de dados, atualmente realizado de maneira manual, é bastante oneroso em termos de tempo e poderia ser otimizado por meio da automação dessa etapa. Além de reduzir o esforço humano, a automação contribuiria para minimizar erros de inserção de dados, garantindo maior confiabilidade no repositório. Uma possibilidade seria o recurso de ingestão por lotes (*batch import*) disponível no *DSpace 7*, que permite a inserção massiva de registros e padroniza o fluxo de ingestão, reduzindo a necessidade de intervenção manual.

4 CONSIDERAÇÕES FINAIS

A aplicação da Inteligência Artificial na organização de bases de dados pode representar um avanço significativo para a otimização de processos na Ciência da Informação, permitindo maior agilidade na estruturação e gestão de repositórios. O uso de ferramentas como o *ChatGPT* possibilita a automação de tarefas operacionais, como a geração de metadados e a padronização de descritores, além de reduzir o tempo necessário para a realização dessas atividades, tornando a administração das informações mais eficiente e acessível. No entanto, é fundamental considerar que, apesar das vantagens proporcionadas pela Inteligência Artificial, a validação e a supervisão humanas continuam indispensáveis para garantir a qualidade, a consistência e a adequação dos dados armazenados.

No caso deste estudo, a base de dados desenvolvida atende às necessidades dos pesquisadores vinculados à Rede Interdisciplinar de Estudos sobre Violências, oferecendo um repositório temático estruturado e de fácil acesso, além de se apresentar como uma ferramenta estratégica para o fortalecimento das pesquisas sobre o tema. Ao reunir e organizar informações de forma padronizada, a base contribui para a disseminação do conhecimento sobre os diferentes tipos de violência, proporcionando uma abordagem mais sistemática na recuperação de dados e na formulação de novas investigações acadêmicas.

Mencionamos ainda que a adoção dos princípios FAIR expande o potencial de colaboração científica, promovendo a interoperabilidade com outras bases e repositórios e incentivando o compartilhamento responsável de informações.

Um desafio encontrado na condução do projeto foi a dificuldade em localizar profissionais na área de Tecnologia da Informação com domínio no processo de configuração do *DSpace 7*. Por conta disso, muitas de nossas demandas referentes à customização desse *software*, no que diz respeito à configuração de *layouts* e à estrutura de apresentação de metadados, foram afetadas.

Este estudo mostra a importância da incorporação de novas tecnologias na gestão de dados, ao mesmo tempo em que evidencia a necessidade de reflexões contínuas sobre as possibilidades e desafios do uso da Inteligência Artificial na Ciência da Informação.

REFERÊNCIAS

- FAIRDATABR. **FairDataBR**: uma ferramenta para a avaliação de conjuntos de dados. Disponível em: <https://wrcu.ufpb.br/fair/index.html>. Acesso em: 24 fev. 2025.
- GABRIEL JUNIOR, R. F.; LAIPELT, R. C. Thesa: ferramenta para construção de tesauro semântico aplicado interoperável. **Revista P2P & Inovação**, Rio de Janeiro, v. 3, n. 2, p. 124-145, mar./set. 2017.
- RUSSEL, S.; NORVIG, P. **Inteligência artificial**. Rio de Janeiro: Elsevier, 2004.
- UNIVERSIDADE FEDERAL DA PARAÍBA. **Rede interdisciplinar de estudos sobre violências**. Disponível em: <https://www.ufpb.br/riev/index.php>. Acesso em: 20 fev. 2025.
- WILKINSON, M. D. *et al.* The FAIR guiding principles for scientific data management and stewardship. **Scientific Data**, [S. l.], v. 3, art. 160018, 2016. Disponível em: <https://www.nature.com/articles/sdata201618>. Acesso em: 20 fev. 2025.

AGRADECIMENTOS

Trabalho apoiado pelas Chamadas FAPESQ-PB/CNPq nº 77/2022 e CNPq nº 09/2018.



[4.0 Internacional](https://creativecommons.org/licenses/by/4.0/)

Esta licença permite compartilhamento, remixe, adaptação e criação a partir do trabalho, mesmo para fins comerciais, desde que sejam atribuídos créditos ao(s) autor(es). Conteúdos elaborados por terceiros, citados e referenciados nesta obra não são cobertos pela licença.

Como citar este trabalho:

DIAS, Guilherme Ataíde; OLIVEIRA, Tassyara Onofre; CAMPELO, Arthur Antônio Martins Sales; CARVALHO, Vivian Roberta Lima de. Utilizando o ChatGPT como ferramenta de suporte para a construção de uma base de dados textual sobre violências: apresentação e análise do processo de desenvolvimento. *In: WORKSHOP DE INFORMAÇÃO DADOS E TECNOLOGIA*, 8., 2025, Marília, SP. **Anais [...]**. Marília, SP: Universidade de Marília, 2025. DOI: <https://doi.org/10.22477/viii.widat.224>.