26º ENANCIB
Encontro Nacional de Pesquisa e Pós-Graduação em Ciência da Informação
A Amazônia na sociedade global em rede: assimetrias e potencialidades informacionais, tecnológicas e ambientais
Belém, Pará, 09 a 13 de novembro de 2026
Universidade Federal do Pará

ENCONTRO NACIONAL DE PESQUISA E PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO
ISSN 2177-3688

GT 8 – Dados, Informação e Tecnologia

DATA WRANGLING, CURADORIA E JUSTIÇA ALGORÍTMICA: TRANSIÇÃO PARADIGMÁTICA PARA UMA INTELIGÊNCIA ARTIFICIAL INCLUSIVA

DATA WRANGLING, CURATION, AND ALGORITHMIC JUSTICE*: A PARADIGM SHIFT TOWARD INCLUSIVE ARTIFICIAL INTELLIGENCE*

Januário Albino Nhacuongue – Universidade Federal de São Carlos (UFSCar)

Modalidade: Trabalho Completo

Resumo: Este trabalho analisa como práticas de curadoria digital e *data wrangling*, articuladas aos paradigmas estruturalista radical e humanista radical da Ciência da Informação (CI), podem sustentar estratégias de coleta e tratamento de dados capazes de promover justiça, transparência e responsabilidade em sistemas de Inteligência Artificial (IA). O referencial teórico integra os conceitos de curadoria digital, regimes de informação e transição paradigmática. Metodologicamente, o estudo combina técnicas de *data wrangling* e Processamento de Linguagem Natural (PLN) aplicadas a um *dataset* de comentários extraídos do canal do Governo Federal (CanalGov) no YouTube. Os resultados indicam que procedimentos como enriquecimento, pseudonimização, normalização lexical e detecção de vieses contribuem para a conversão de dados brutos em ativos informacionais alinhados aos princípios *Findable, Accessible, Interoperable e Reusable (FAIR) e Fairness, Accountability, Transparency e Ethics (FATE)*. Conclui-se que o enfrentamento da opacidade algorítmica requer práticas de curadoria orientadas pela diversidade, reflexividade e governança ética dos dados. Esse desafio impõe uma reconfiguração paradigmática, de caráter holístico e crítico, que reposicione a Ciência da Informação na arena estratégica para o desenvolvimento de sistemas de IA mais transparentes, responsáveis e socialmente justos.
Palavras-chave: curadoria digital; ciclo de vida dos dados; regimes de informação; data wrangling; inteligência artificial.

Abstract: This study analyzes how digital curation and data wrangling practices, linked to the radical structuralist and radical humanist paradigms of Information Science (IS), can support data collection and processing strategies capable of promoting justice, transparency, and

accountability in Artificial Intelligence (AI) systems. The theoretical framework integrates the concepts of digital curation, information regimes, and paradigmatic transition. Methodologically, the study combines data wrangling and Natural Language Processing (NLP) techniques applied to a dataset of comments extracted from the Federal Government's YouTube channel (CanalGov). The results indicate that procedures such as enrichment, pseudonymization, lexical normalization, and bias detection contribute to the conversion of raw data into informational assets aligned with the Findable, Accessible, Interoperable, and Reusable (FAIR) and Fairness, Accountability, Transparency, and Ethics (FATE) principles. We conclude that addressing algorithmic opacity requires curatorial practices guided by diversity, reflexivity, and ethical data governance. This challenge calls for a holistic and critical paradigm shift that repositions Information Science in the strategic arena for the development of more transparent, accountable, and socially just AI systems.
Keywords: digital curation; data lifecycle; information regimes; data wrangling; artificial intelligence.

1 INTRODUÇÃO

Desde sua emergência no final do século XX, o informacionalismo, conceito formulado por Castells (2025), vem redefinindo uma nova ordem sociotécnica, estruturada em torno da produção e uso de conhecimento, do processamento intenso da informação e da comunicação. Estruturado por uma lógica de redes, esse paradigma não substitui o industrialismo, mas o reconfigura, tornando a informação um recurso produtivo chave e organizando funções sociais em fluxos operados em tempo real.

No cenário global atual, o informacionalismo manifesta-se por meio de uma profunda transformação digital que transcende a mera atualização técnica, redefinindo as relações econômicas, sociais e institucionais. Dados do relatório *Global Digitalization and Intelligence Index 2025 (GDII)*[1] estimavam que a economia digital global deveria atingir a marca de US$ 26,7 trilhões até o final de 2025, representando cerca de 25% do PIB mundial. Essa evolução impõe às instituições uma transformação estratégica profunda em toda a sua cadeia de valor.

Para Rogers (2017), a mudança estratégica de organizações demanda uma visão holística, fundamentada na redefinição de cinco domínios chaves: usuários, competição, dados, inovação e valor. O percurso dessa mudança, no Brasil, é orientado por diversos documentos, com destaque para a Estratégia Brasileira para a Transformação Digital (E-Digital 2022-2026),

[1] Elaborado pela Huawei em parceria com a empresa *International Data Corporation (IDC)*, oferece um panorama abrangente sobre a evolução digital das economias no contexto da transformação impulsionada pela Inteligência Artificial (IA). Estruturado em sete pilares (criação de dados, transferência de dados, processamento e armazenamento de dados, aplicação de dados, energia digital, políticas, talento e ecossistema), o estudo analisa 90 economias que representam 94% do PIB global, fornecendo evidências e métricas para orientar políticas, investimentos e estratégias de desenvolvimento.

que busca coordenar ações em diferentes eixos para elevar a produtividade e a competitividade nacional. O Modelo de Maturidade de Dados (MMD) atua como documento de referência na avaliação de níveis organizacionais do setor público brasileiro, a partir da sua capacidade de gerenciar, governar, proteger e utilizar seus dados para inovação e reinvenção da sua cadeia de valor (Ministério da Gestão e Inovação em Serviços Públicos, 2026). A Estratégia Brasileira de Inteligência Artificial (EBIA) e o Plano Brasileiro de Inteligência Artificial (PBIA 2024-2028), por sua vez, estabelecem os fundamentos necessários para o desenvolvimento e a aplicação ética da IA, articulando-se aos temas estratégicos da Estratégia Nacional de Ciência, Tecnologia e Inovação (ENCTI 2024-2034) e dos observatórios do Centro de Gestão e Estudos Estratégicos (CGEE).

Na Ciência da Informação, as Humanidades Digitais surgem como um campo interdisciplinar que busca compreender os impactos sociais e culturais dessa nova realidade. A relação simbiótica estabelecida entre Ciência de Dados e Humanidades Digitais representa um desafio urgente, tanto pelas novas competências exigidas dos profissionais da informação quanto pelos novos contextos de atuação voltados à geração de valor baseado em dados (Paletta; Silva; Mucheroni, 2021). Esse desafio ganha contornos mais amplos quando analisado à luz de diretrizes macro de instituições como a Federação Internacional de Associações e Instituições Bibliotecárias (IFLA).

A Declaração sobre Bibliotecas e IA e seus anexos, da IFLA (2020), estabelecem diretrizes sobre as implicações da IA na promoção da liberdade intelectual, alertando para riscos associados a *datasets* desbalanceados, filtros opacos e processos automatizados que podem restringir a liberdade e diversidade em ambientes digitais. Nesse contexto, aplicações emergentes de PLN, como a classificação automática de coleções por meio da extração de entidades e relações, podem ampliar vieses e reproduzir estereótipos presentes nos dados de treinamento. Tal cenário reforça a necessidade de fortalecimento da literacia em IA.

As declarações da IFLA (2020), por um lado, alinham-se ao modelo de ciclo de vida de dados e seus fatores transversais associados, como privacidade, qualidade, integração, direitos autorais, disseminação e preservação, proposto por Santana (2016), e ao conceito de curadoria ativa, proposto pela iniciativa *Digital Curation Centre - DCC* (DeRidder, 2018). Por outro lado, demonstram interseção com o conceito de regime de informação na CI, compreendido como o modo informacional dominante em uma formação social que estipula quem são os sujeitos, as regras, as autoridades e os recursos preferenciais de informação em determinado tempo e

espaço. Ou seja, diferentes arranjos sociotécnicos moldam quem detém autoridade informacional e como o poder opera na definição de padrões e na circulação do conhecimento (González de Gómez, 2012, 2017).

O presente trabalho surge e justifica-se na tessitura dos desafios da transformação digital apresentados, mormente os processos de curadoria de dados e *data wrangling* ou conjunto de atividades destinadas a transformar e preparar dados brutos em formatos refinados prontos para análise (Ratner; Kasakoff, 2017). Nesse contexto, busca-se respostas a questões que se tornam cada vez mais urgentes na CI: que paradigmas e estratégias de coleta e tratamento de dados são essenciais para sustentar o desenvolvimento de uma IA justa, transparente e socialmente responsável, capaz de mitigar a reprodução de estereótipos e a polarização de saberes no campo tensionado entre a intersubjetividade das práticas locais de informação e as demandas de padronização de infraestruturas globais?

Embora a resposta a estes questionamentos não seja trivial e nem possa ser construída a partir de uma única perspectiva, o objetivo de analisar como práticas de curadoria digital e *data wrangling*, articuladas aos paradigmas estruturalista radical e humanista radical da CI, podem sustentar estratégias de coleta e tratamento de dados capazes de promover justiça, transparência e responsabilidade no desenvolvimento de sistemas de IA, evidencia progressões teóricas e empíricas de relevo. Esses avanços permitem compreender, na prática, o afastamento da restritiva tríade dos paradigmas físico, cognitivo e social em direção ao arcabouço holístico e crítico, fornecido pelos paradigmas estruturalista radical e humanista radical, conforme proposto por Araujo (2021).

Desenvolvido no âmbito de uma disciplina optativa do curso de Biblioteconomia e Ciência da Informação da UFSCar, este estudo teve caráter eminentemente formativo, visando introduzir os estudantes aos desafios da transformação digital e às competências relacionadas aos dados. Em consonância com esse propósito, foram utilizados *datasets* de menor complexidade e procedimentos analíticos introdutórios, priorizando a compreensão dos fundamentos metodológicos em detrimento de análises computacionais mais avançadas.

2 CURADORIA DIGITAL, CICLO DE VIDA DE DADOS E DATA WRANGLING: CONVERGÊNCIAS EPISTEMOLÓGICAS E TENSÕES NO CAMPO INFORMACIONAL

A consolidação da curadoria digital como disciplina, atualmente reconhecida como um campo complexo e dinâmico que vai além da preservação, coincide com as mudanças do

ecossistema informacional para um cenário em que dados, metadados e processos computacionais se tornam indissociáveis. A literatura demonstra que a curadoria digital nasce na CI, mas rapidamente se expande para um território híbrido, no qual fundamentos arquivísticos e biblioteconômicos, métodos computacionais e modelos de gestão de dados convergem para sustentar a perenidade da informação digital ao longo do tempo. Higgins (2018) observa que a curadoria digital se estrutura em torno de objetos digitais e de um paradigma unificador, orientado a garantir seu uso contínuo. Por isso, embora tenha emergido tanto da tradição arquivística, voltada à preservação de documentos digitais nativos e à manutenção de sua integridade, quanto da Biblioteconomia, especialmente em relação às práticas de digitalização de materiais analógicos para preservação, sua abrangência atual atravessa todas as subáreas da CI como uma sub-metadisciplina.

A centralidade do dado como entidade viva, sujeita a transformações, deslocamentos e reinterpretações, aproxima a curadoria digital das discussões sobre modelos de ciclo de vida dos dados. Sant'Ana (2016) propõe um modelo em que os dados constituem o eixo organizador das práticas, e não meros subprodutos de sistemas ou processos. As quatro fases (coleta, armazenamento, recuperação e descarte) e seus fatores transversais (privacidade, integração, qualidade, direitos autorais, disseminação e preservação) orientam esse eixo.

Os fatores transversais associados ao modelo de ciclo de vida, apontados por Sant'Ana (2016), traduzem a essência dos desafios das Humanidades digitais na CI, que abrangem dimensões epistemológicas, institucionais e técnicas interconectadas. Do ponto de vista crítico e político, por exemplo, torna-se necessário superar a dependência em relação ao Norte Global, combatendo o neocolonialismo de dados e os diferentes tipos de vieses embutidos em modelos globais por meio de abordagens decoloniais, interseccionais e multilíngues (Josiowicz; Kothari, 2023; Paletta; SIlva; Mucheroni, 2021).

Nesse contexto, a transformação digital (Rogers, 2017) na CI demanda uma reconfiguração de seus domínios estratégicos. Da concepção de rede de consumidores, o domínio "usuários" evolui para sujeitos de direitos, cuja privacidade e acessibilidade devem ser asseguradas. O domínio "competição" é ressignificado pelos princípios da Ciência Aberta, que orientam a interoperabilidade e integração entre ecossistemas de dados, em contraste com a opacidade de ambientes proprietários. O domínio "dados" deixa de ser um simples recurso de inteligência para assumir papel central, demandando curadoria ativa e documentação alinhada aos princípios FAIR e FATE. O domínio "inovação" desloca seu foco da velocidade de

experimentação orientada pelo mercado para incorporar práticas de inovação aberta. Por fim, o domínio "valor" deixa de ser medido apenas por métricas financeiras ou de retenção para ser compreendido como impacto social e valor público, com ênfase na redução de assimetrias informacionais e no compromisso ético com a sociedade.

Nesse ponto, a interseção entre o modelo de ciclo de vida de dados na CI e o campo do *data wrangling*, entendido como o conjunto de práticas que viabilizam a transformação de dados brutos em dados analiticamente úteis, torna-se marcante. Rattenbury *et al.* (2017) afirmam que entre 50% e 80% do tempo de trabalho analítico é consumido por atividades de preparação, limpeza, estruturação e enriquecimento dos dados, constituindo o cerne do processo profissional. O *data wrangling* opera, portanto, na zona liminar entre o dado bruto e o dado interpretável, desempenhando papel equivalente ao da curadoria digital no que diz respeito à mediação entre materialidade técnica e valor informacional.

Para Davenport e Kim (2014), sem processos consistentes de tratamento, organizações desperdiçam dados, na medida em que a analítica (descritiva, preditiva ou prescritiva), depende de dados limpos, estruturados e contextualizados, o que coloca o *data wrangling* como fluxo indispensável para transformar dados em valor.

A aproximação entre curadoria digital, ciclo de vida de dados e *data wrangling* não é meramente terminológica. Pelo contrário, ela revela uma isomorfia conceitual que traduz a essência do problema: como garantir que dados continuem significativos, utilizáveis e confiáveis ao longo do tempo? E, mais do que isso, como garantir que o próprio processo de modelagem, responsável por erguer o arcabouço lógico que sustenta tanto a interpretação quanto a geração, permaneça imune às distorções sutis, porém decisivas, presentes nos dados de treinamento?

Na convergência entre curadoria digital, ciclo de vida de dados e *data wrangling*, três aspectos se destacam. O primeiro diz respeito à centralidade do dado como objeto epistemológico. Tanto Higgins (2018) quanto Davenport e Kim (2014) e Sant'Ana (2016) deslocam o foco das infraestruturas tecnológicas para os dados em si. O segundo é referente à temporalidade como eixo organizador. O ciclo de vida dos dados fornece a arquitetura temporal e estrutural; a curadoria digital estabelece o quadro de governança e preservação de valor; e o *data wrangling* opera como a engrenagem iterativa de refinamento que transforma e qualifica os dados. O terceiro diz respeito à mediação humano-tecnológica, palco onde se articulam disputas de poder inerentes aos regimes de informação. Esse cenário evidencia a necessidade

de uma transição paradigmática capaz de reposicionar a CI diante das urgências decoloniais e das tensões estruturais que atravessam a sociedade em rede.

2.1 REGIMES DE INFORMAÇÃO E A NECESSÁRIA TRANSIÇÃO PARADIGMÁTICA NA CIÊNCIA DA INFORMAÇÃO

O entendimento de que a técnica e o humano coexistem em uma relação de codeterminação obriga a área da CI a confrontar o conceito de regime de informação. Segundo González de Gómez (2012, 2017), tal constructo designa o modo informacional dominante que estipula, em determinado tempo e lugar, quem são os sujeitos autorizados, quais as regras de validação e quais recursos de informação possuem prioridade. Nesse contexto, a curadoria deixa de ser compreendida como um conjunto de procedimentos estritamente técnicos para assumir o caráter de prática sociopolítica inserida em um plexo de relações de poder. A existência de um arquivo tecnológico global não garante a equidade; ao contrário, diferentes regimes podem constituir-se a partir de infraestruturas semelhantes, resultando tanto em modelos monopólicos que centralizam a informação quanto em regimes abertos voltados para a preservação de memórias coletivas e acesso universal.

A noção de regime de informação, como propõe González de Gómez (2012), ganha contornos críticos quando confrontada com a realidade da opacidade algorítmica descrita por Pasquale (2015). O autor alerta que vivemos em uma “sociedade de caixas-pretas”, onde sistemas proprietários e secretos tomam decisões estratégicas sobre reputação, finanças e acesso a recursos sem que os indivíduos possam contestar ou sequer compreender a lógica envolvida. Essa assimetria de poder estabelece um espelho de face única, no qual corporações possuem um conhecimento granular da vida privada dos cidadãos, enquanto seus próprios métodos permanecem protegidos por camadas de segredo comercial e complexidade técnica.

As evidências apresentadas por Noble (2018) reforçam essa compreensão ao demonstrar que os sistemas algorítmicos não são infraestruturas neutras, mas mecanismos capazes de reproduzir e amplificar desigualdades sociais. A autora evidencia como motores de busca podem privilegiar resultados marcados por estereótipos raciais e de gênero, notadamente em relação a mulheres negras e grupos historicamente marginalizados, em função de lógicas econômicas que convertem visibilidade em valor mercadológico. Tal fenômeno revela a dimensão política da organização da informação e escancara a indecibilidade

estrutural[2] da informação, frequentemente associada a critérios institucionais, tecnológicos e econômicos que favorecem interesses hegemônicos. Nessa perspectiva, conforme argumenta Rabello (2018), a informação não possui significado ou estatuto ontológico fixos, adquirindo estabilidade apenas por meio de processos de mediação e institucionalização. Assim, a definição do que será reconhecido, recuperado e legitimado como informação resulta de disputas simbólicas contínuas.

O cenário analisado reforça a necessidade, apontada por Araújo (2024), de transcender a tríade clássica dos paradigmas físico, cognitivo e social que historicamente estruturou a CI. O fulcro dessa transição reside no deslocamento de abordagens predominantemente funcionalistas e interpretativistas para perspectivas críticas, capazes de ampliar o escopo analítico e revelar as estruturas de poder, os condicionantes sociotécnicos e as assimetrias que permeiam os fenômenos informacionais.

Para Araújo (2024), os modelos tradicionais físico, centrado na informação como sinal tangível, e o cognitivo, focado nas estruturas mentais individuais, são pertencentes primordialmente aos quadrantes funcionalista e interpretativo. Uma transição verdadeiramente profunda, segundo o autor, exige a inclusão do paradigma estruturalista radical e do paradigma humanista radical. No paradigma estruturalista radical, a informação é compreendida como parte constituinte das disparidades estruturais de classe, gênero e raça, o que torna a curadoria ativa um instrumento essencial para denunciar a opacidade algorítmica e as caixas-pretas que governam o fluxo de dados global. Paralelamente, o paradigma humanista radical oferece o senso de direção para a emancipação do sujeito, transformando a literacia em IA e a competência em informação em ferramentas de resistência contra a alienação digital e o colonialismo de dados. Assim, a transição para um paradigma social pleno não deve ser vista apenas como uma nova categoria, mas como uma abertura epistemológica que legitima visões críticas e decoloniais antes invisibilizadas. O objetivo dessa redefinição paradigmática é garantir que a CI compreenda a informação não apenas como um objeto físico ou mental, mas como um instrumento de poder e de transformação social.

[2] A indecibilidade estrutural designa a condição pela qual a informação permanece aberta a múltiplas possibilidades de significação antes da intervenção de processos seletivos e decisórios. Sob essa perspectiva, a informação não possui uma estrutura intrínseca ou previamente determinada; ao contrário, seus significados e valores são produzidos por processos de sobredeterminação decorrentes de escolhas, explícitas ou implícitas, realizadas por indivíduos e grupos sociais em contextos específicos de ação e tomada de decisão (Rabello, 2018).

No tocante ao processo de curadoria de dados, a transição paradigmática crítica na CI se ancora em documentos estratégicos globais, como os princípios *Fairness, Accountability, Transparency e Ethics (FATE)* ou Justiça, Responsabilidade, Transparência e Ética. Esses princípios buscam harmonizar os sistemas de IA, promovendo transparência e responsabilidade por meio de práticas documentadas e auditáveis, que sejam capazes de prevenir vieses nos processos de inferência. Em síntese, a construção de um regime de informação pleno requer uma CI capaz e disposta a assumir seu papel político na denúncia da opacidade estrutural e na promoção de uma infraestrutura de dados orientado ao interesse público, conforme preconizado no Plano Brasileiro de Inteligência Artificial.

3 PROCEDIMENTOS METODOLÓGICOS

Os fundamentos metodológicos desta pesquisa ancoram-se nas sete etapas de procedimento propostas por Quivy e Campenhoudt (2019). Esse percurso inicia-se com a formulação da pergunta de partida e o subsequente processo de exploração, evolui para a definição da problemática e a construção do modelo de análise, e culmina nos estágios de observação, análise e conclusões.

O modelo analítico integra o conceito de ciclo de vida dos dados sob a perspectiva da CI, estruturado nas fases de coleta, armazenamento, recuperação e descarte. Fatores transversais críticos, tais como privacidade e qualidade, acompanham cada estágio desse fluxo informacional (Sant'Ana, 2016).

Conforme destacado na introdução, este estudo resulta de uma disciplina optativa ministrada no curso de Bacharelado em Biblioteconomia e Ciência da Informação da UFSCar, cujo objetivo foi inserir os estudantes nas discussões e práticas relacionadas à transformação digital e às novas competências profissionais requeridas nesse cenário. Em função desse propósito, optou-se por utilizar *datasets* de menor escala, o que limita seu escopo analítico a cenários de menor complexidade e tensionamento informacional.

O *dataset* analisado (*dataset_youtube-comments-scraper*) foi obtido a partir do canal oficial do Governo Federal brasileiro no YouTube (*CanalGov*). A coleta de dados *(web scraping)*, ocorreu em duas etapas complementares. Na primeira, empregou-se o *actor* "*YouTube Scraper"* na plataforma *Apify (https://apify.com/)*, com o objetivo de extrair metadados dos vídeos publicados no canal. Após a exportação dos registros em formato CSV, procedeu-se à segunda etapa, utilizando o *actor* "*YouTube Comments Scraper*" para coletar os comentários associados

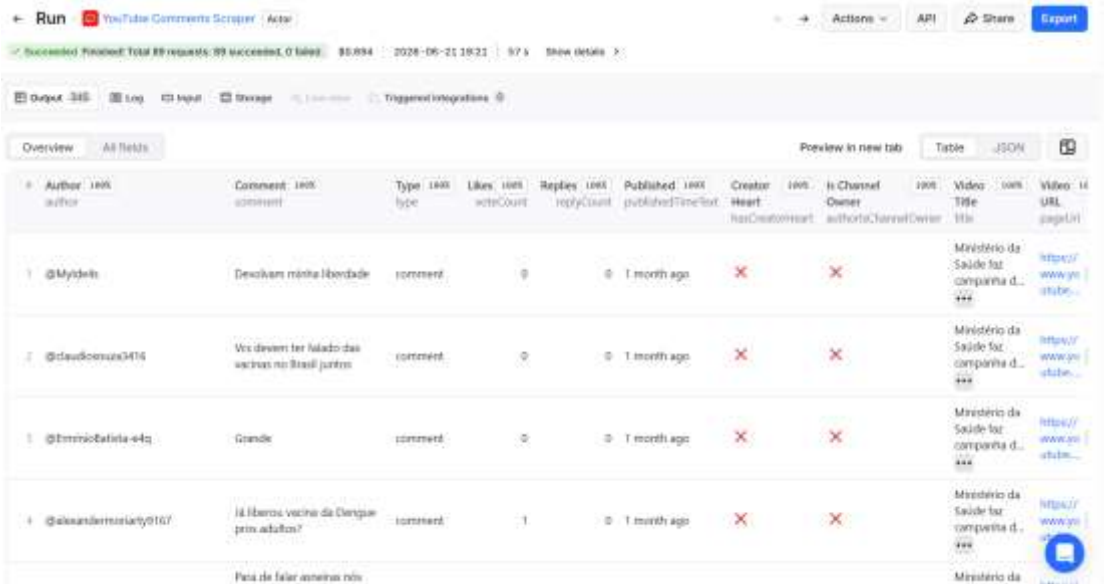
aos vídeos selecionados, que constituíram a principal unidade de análise deste estudo. Os dados foram extraídos em abril de 2026, considerando como critérios de inclusão os vídeos mais recentemente publicados no canal e seus respectivos comentários.

O tratamento dos dados linguísticos brutos provenientes do *dataset* do CanalGov foi conduzido por meio das etapas de *data wrangling* (*raw stage, profiling, structuring, cleansing, enriching, validation e publishing*). Esse processo buscou estruturar, qualificar e validar os dados, tornando-os adequados às etapas subsequentes de análise (Rattenbury *et al.*2017). No contexto do Processamento de Linguagem Natural (PLN), a preparação do *corpus* textual envolveu uma sequência de operações, como tokenização para segmentação do texto em unidades léxicas e a normalização linguística, incluindo a conversão para caracteres minúsculos, a lematização e a radicalização (*stemming*), com vistas à redução da variabilidade morfológica e ao aumento da consistência dos dados (Caseli; Nunes, 2024). Do ponto de vista ético e do paradigma social, o estudo se fortalece pela aplicação de técnicas de anonimização e pseudonimização, protegendo atributos sensíveis dos usuários comentaristas em conformidade com os princípios FATE (Pavão *et al.* 2026).

4 RESULTADOS E DISCUSSÃO

No fluxo de *data wrangling*, a etapa inicial, denominada *raw stage*, é dedicada à exploração inicial dos dados e à produção de metadados essenciais para sua interpretação, contextualização e gestão ao longo do ciclo de vida dos dados enquanto ativo estratégico. Essa fase estrutura-se em torno de três ações principais: a ingestão propriamente dita, a geração de metadados genéricos e a definição de metadados proprietários (Rattenbury *et al.*2017).

Figura 1 - Etapa de ingestão de dados *(raw stage)*



Fonte: elaboração própria (2026).

No *dataset_youtube-comments-scraper*, utilizado como estudo de caso, a unidade de granularidade corresponde aos vídeos publicados no canal CanalGov. O *dataset* é composto por metadados descritivos, como *author* (autor do comentário), *viewCount* (número de visualizações), *date* (data de publicação) e *title* (título do vídeo), entre outros.

Um aspecto importante na articulação entre o fluxo de *data wrangling* e os princípios FAIR e FATE consiste na incorporação de mecanismos que favorecem a rastreabilidade, a transparência e a reprodutibilidade dos processos. No *dataset* analisado, elementos como o *Run ID (fShLoEcytjh8yVcHQ)* ou identificador único da execução desempenham papel chave nesse processo. A sequência de caracteres funciona como um *token* de autenticação para recuperar detalhes de execução como parâmetros de entrada, registros (*logs*) e eventuais interrupções. Da mesma forma, metadados temporais associados à coleta, como os registros de início e término da execução - *Timestamp* (23/04/2026 11:32:21 - 23/04/2026 11:34:56), bem como a versão do *scraper* utilizado (*Streamers/youtube-scraper* 0.0.247), contribuem para a encontrabilidade, a rastreabilidade e a auditabilidade de todo o fluxo de processamento. Quando incorporados à documentação do *dataset*, por exemplo em arquivos *README*, esses elementos ampliam seu potencial de reuso. Em projetos de larga escala, o *Run ID* pode ainda ser integrado a mecanismos como *checksums*, fortalecendo o controle de integridade ao longo das etapas de coleta, transformação e preservação dos dados.

Figura 2 - Etapa de *profiling*

```
Tipos de dados por coluna:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 344 entries, 0 to 343
Data columns (total 14 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   author                344 non-null    object
 1   authorIsChannelOwner  344 non-null    bool
 2   cid                   344 non-null    object
 3   comment               344 non-null    object
 4   commentsCount         344 non-null    int64
 5   hasCreatorHeart       344 non-null    bool
 6   pageUrl               344 non-null    object
 7   publishedTimeText     344 non-null    object
 8   replyCount            344 non-null    int64
 9   replyToCid            44 non-null     object
 10  title                 344 non-null    object
 11  type                  344 non-null    object
 12  videoId               344 non-null    object
 13  voteCount             344 non-null    int64
dtypes: bool(2), int64(3), object(9)

Valores ausentes por coluna:
author                    0
authorIsChannelOwner      0
cid                       0
comment                   0
commentsCount             0
hasCreatorHeart           0
pageUrl                   0
publishedTimeText         0
replyCount                0
replyToCid              300
title                     0
type                      0
videoId                   0
voteCount                 0
dtype: int64
```

Fonte: elaboração própria (2026).

O *profiling*, segunda etapa do fluxo de *data wrangling*, corresponde ao diagnóstico inicial do *dataset*, permitindo avaliar sua estrutura, distribuição de valores e potenciais problemas de qualidade antes das etapas de transformação e análise. Esse processo inclui a verificação de restrições sintáticas e semânticas, bem como a identificação de padrões e anomalias nos dados. No *dataset* analisado, o *profiling* revelou a presença de *outliers* ou valores atípicos, destacando-se o vídeo sobre o fim da escala 6x1, que apresentou cerca de 1.000 comentários, volume significativamente superior ao observado nas demais publicações do canal (Rattenbury *et al*., 2017).

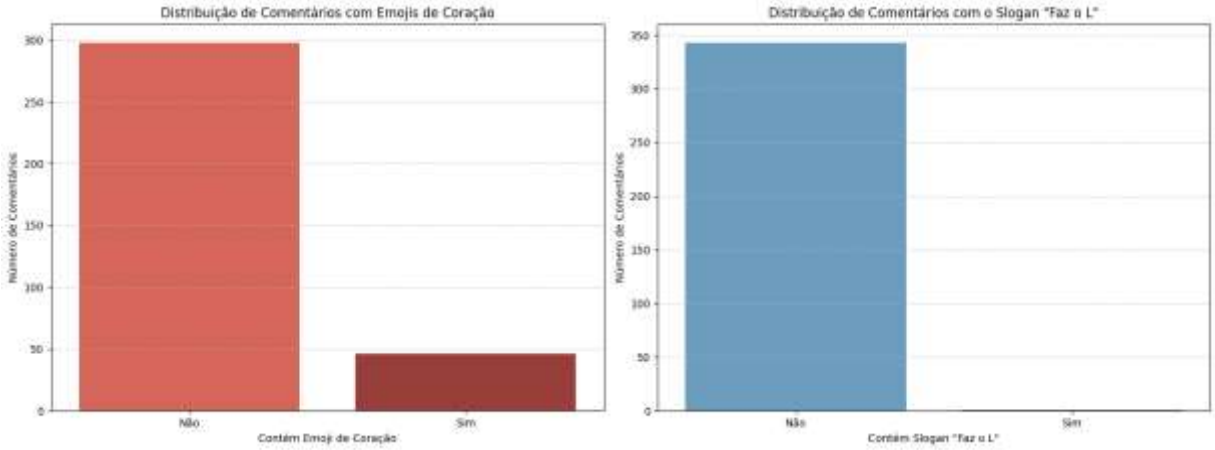
Ao ser carregado em um *DataFrame* da biblioteca *pandas*, o *dataset* é submetido a uma exploração inicial para caracterização da sua estrutura, dos tipos de dados, da distribuição de valores ausentes e de estatísticas descritivas básicas. No *dataset_youtube-comments-scraper*, essa análise revelou a presença de valores ausentes (*NaN*) na coluna *replyToCid*, atributo que identifica o comentário ao qual uma determinada resposta está associada (Figura 2). A presença desses valores demanda tratamento adequado, uma vez que pode introduzir vieses ou comprometer o desempenho do modelo treinado.

A etapa de *data cleansing* no fluxo de *data wrangling* compreende um conjunto de ações voltadas à identificação e correção de inconsistências que possam comprometer a integridade, a qualidade e a confiabilidade dos dados. Essas operações incluem o tratamento de valores ausentes, nulos ou inconsistentes, bem como a eliminação de registros duplicados, mitigando vieses e distorções estatísticas decorrentes de redundâncias. Em aplicações de PLN, a limpeza também envolve a remoção de ruídos textuais, como marcações HTML, URLs e caracteres

especiais, além da normalização linguística, incluindo padronização ortográfica e tratamento de sinais diacríticos (Rattenbury *et al*., 2017; Caseli; Nunes, 2024).

A etapa de *structuring* consiste na reorganização da representação dos dados visando ampliar sua adequação analítica. Esse processo envolve a transformação de atributos e a derivação de novas variáveis a partir de informações já existentes no *dataset*. Por exemplo, a decomposição da *string* complexa "2026-04-16T15:19:53.000Z" em componentes/novas colunas como ano (2026) ou mês (04) permite análises estratificadas. Complementarmente, a etapa de *enriching* corresponde à incorporação de atributos derivados. Nos *datasets* utilizados, por exemplo, podem ser adicionados indicadores como a taxa de engajamento, calculada pela razão entre o número de curtidas (*likes)* e a quantidade de visualizações *(viewCount*). Essa taxa permite inferências mais detalhadas ao longo das análises. No vídeo sobre o fim da escala 6x1, seria de 16,9% (169/999).

Gráficos 1 e 2 - Processamento de Linguagem Natural (etapa *cleasing*)



Fonte: elaboração própria (2026).

O processamento em campos textuais do *dataset_youtube-comments-scraper* incluiu as etapas de tokenização, remoção de *stopwords*, lematização/*stemming*, padronização para letras minúsculas e eliminação de emojis, caracteres especiais e outros elementos ruidosos da coluna *comment*. Os Gráficos 1 e 2 apresentam a distribuição dos comentários que contêm emojis de coração "*has_heart_emoji*" e aqueles que utilizam o slogan "Faz o L" (*has_faz_o_L_slogan*). As colunas (*has_heart_emoji e has_faz_o_L_slogan)* foram adicionadas na etapa *enriching* ou enriquecimento do *data wrangling*. Essa etapa caracteriza-se pela adição deliberada de novos valores e variáveis que não estavam presentes no *dataset* original, expandindo proporcionalmente o seu escopo analítico (Rattenbury *et al.* 2017).

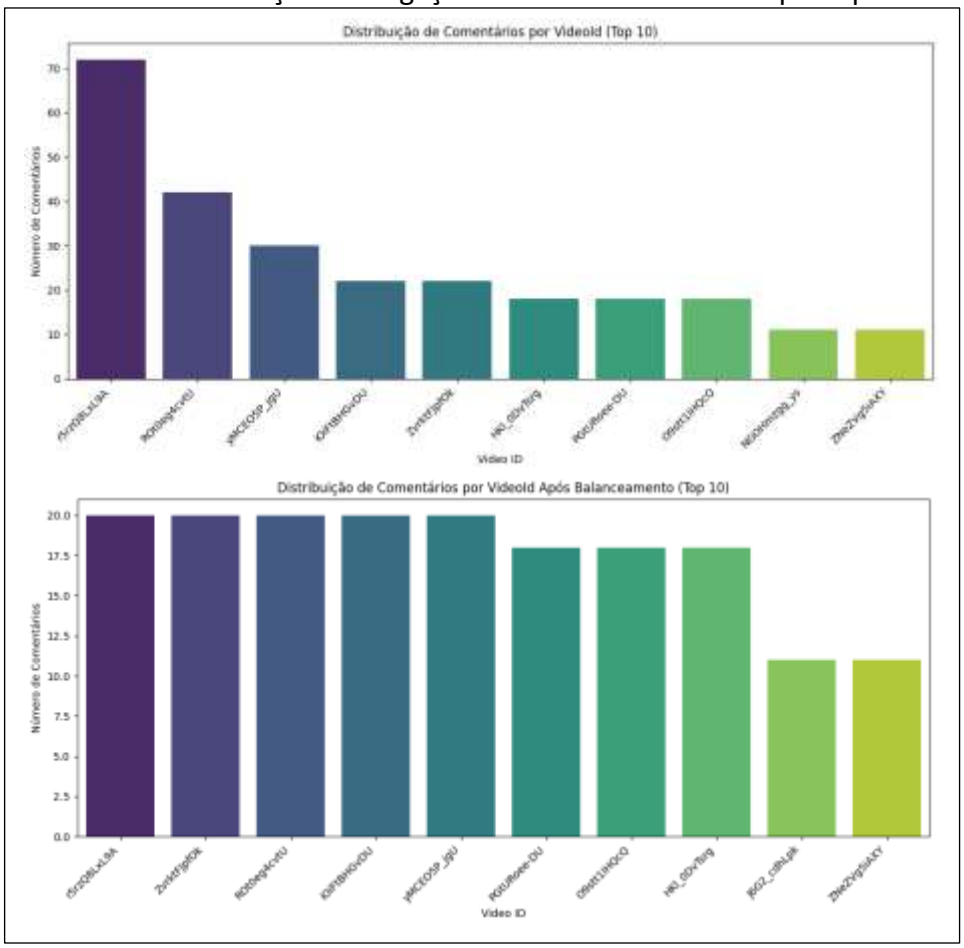
A barra correspondente a "sim" para "*has_heart_emoji*" representa a quantidade de comentários do CanalGov que expressam apoio, afeto ou aprovação através de emoji de coração. Frequências elevadas desse marcador podem indicar maior engajamento emocional positivo em torno de temas específicos. De modo análogo, a barra "sim" para "*has_faz_o_L_slogan*" quantifica comentários que fazem referência a esse *slogan* político. Em uma perspectiva de larga escala, sua incidência permitiria captar dinâmicas de alinhamento político e polarização discursiva, seja na forma de endosso, contestação ou crítica. O processo de enriquecimento incluiu ainda métricas estruturais, como o comprimento do comentário (*comment_length*) e o número de palavras únicas (*unique_words_count*) para análises futuras.

A integração dos princípios FATE foi feita por meio estratégias voltadas à proteção da privacidade, mitigação de vieses e padronização dos dados. Assim, a pseudonimização da variável *"author"* mediante a aplicação de *hashes* criptográficos impediu a associação direta entre comentários e indivíduos específicos, reduzindo riscos de exposição, discriminação e uso indevido de informações pessoais (Memarian; Doleck, 2023; Noble, 2018). A convergência de diferentes formas lexicais ("vacina", "vacinas" e "vacinação") para "vacina", garantiu que todas as menções ao tema tivessem o mesmo peso na análise.

A detecção de vieses buscou identificar e mitigar distorções relacionadas a representações demográficas implícitas nos comentários, bem como padrões linguísticos capazes de indicar formas sutis de preconceito. O gráfico superior "Distribuição de comentários por *VideoId* (Top 10)" é um bom exemplo para compreender a heterogeneidade do *dataset_youtube-comments-scraper.* Há uma clara concentração de um grande volume de comentários em um número limitado de *videoIds*. O *videoId* "r5rzQ8LxL9A", por exemplo, domina significativamente, com 72 comentários, seguido por "ROt0eg4cvtU" com 42 comentários e "yMCEO5P_jgU" com 30 comentários. Isso indica que a discussão sobre o tema da vacinação não está distribuída uniformemente pelos vídeos, mas sim polarizada em torno de alguns conteúdos específicos. Essa distribuição desigual é um forte indicativo de um desequilíbrio de dados[3], que, em uma escala maior, poderia afetar modelos de inferência desenvolvidos a partir do *dataset*.

[3] Se a análise fosse realizada com o *dataset* não balanceado, os *videoIds* com maior número de comentários poderiam super-representar certas narrativas, opiniões ou tópicos, enviesando as conclusões gerais sobre o discurso público.

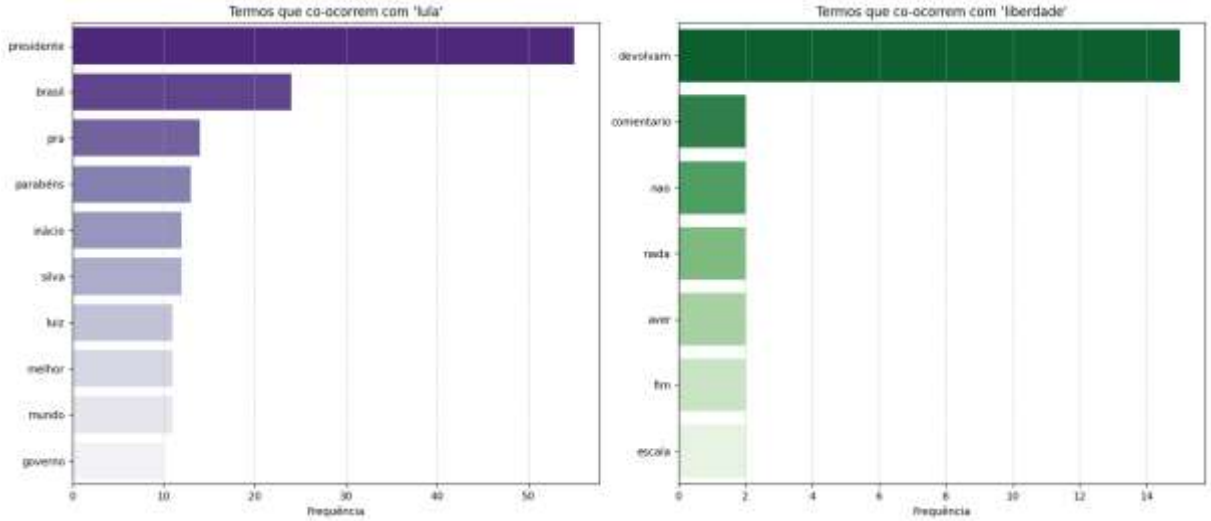
Gráficos 3 e 4 - Detecção e mitigação de vieses conforme os princípios FATE



Fonte: elaboração própria (2026).

O gráfico inferior, intitulado “Distribuição de comentários por *VideoId* após balanceamento (Top 10)”, evidencia os efeitos da aplicação da técnica de *downsampling* no *dataset*. Ao limitar o número de comentários por vídeo a um máximo de 20, reduz-se a influência desproporcional de conteúdos com elevado engajamento. Como resultado, o *dataset* passa a representar de forma mais equilibrada a diversidade temática e discursiva presente no *corpus*, evitando que *hotspots* de engajamento dominem as análises e comprometam a interpretação dos padrões observados. Outro ponto importante de balanceamento de dados recai sobre regimes de informação, cujo conceito foi abordado anteriormente. Através das etapas de PLN, normalização de linguagem e análises de frequência e coocorrência de termos, foi possível observar a delineação desses regimes no *dataset_youtube-comments-scraper.*

Gráficos 5 e 6 - Detecção de vieses: análise de co-ocorrência



Fonte: elaboração própria (2026).

A identificação das palavras mais frequentes, como “Lula”, “presidente”, “Brasil”, “governo”, “vacina” e “liberdade” (Gráficos 5 e 6), aponta para pilares temáticos que estruturam os principais regimes de informação. Estes termos atuam como eixos em torno dos quais as discussões se agrupam. A forte coocorrência de “Lula” com “presidente” e “governo” sugere um regime de informação centrado em discussões sobre a figura presidencial e políticas do governo. Dentro desse regime, termos como “parabéns”, “melhor” e “vida” coocorrem com “Lula”, indicando uma narrativa de apoio, enquanto outros como “devolvam” podem ser associados a narrativas de crítica. Esses padrões lexicais e suas associações não apenas identificam os temas centrais, mas também revelam nuances de narrativas, sentimentos predominantes e ideologias presentes nos comentários do CanalGov. Em dimensões analíticas mais profundas, compreender esses regimes discursivos é essencial para examinar a polarização, a desinformação e as dinâmicas de engajamento no discurso *online*. Embora este estudo não avance nessas camadas investigativas, oferece trilhas para investigações futuras e fortalecimento de competências em dados como eixo estratégico da CI.

O *dataset* processado foi gerado em formato aberto CSV e enriquecido por um arquivo de metadados abrangente que descreve o *dataset*. Esse documento detalha, de forma transparente e auditável, todas as estratégias adotadas para mitigar vieses, promover equilíbrio amostral e assegurar rastreabilidade completa dos dados, articulando-se às diretrizes do *Aether Data Documentation* da Microsoft (2022), referência complementar aos princípios FATE.

5 CONSIDERAÇÕES FINAIS

A investigação realizada demonstra que a Ciência da Informação possui o arcabouço teórico e técnico necessários para confrontar a hegemonia de sistemas automatizados opacos. A transição para os paradigmas estruturalista radical e humanista radical, conforme defendido por Araújo (2024), torna-se necessária para desconstruir a ilusão de neutralidade técnica e evidenciar como os algoritmos podem atuar como dispositivos de exclusão. Os resultados da aplicação prática de princípios FATE ao *dataset* do governo federal confirmam que a curadoria ativa e o *data wrangling* criterioso constituem a primeira linha de defesa contra a reprodução de preconceitos sistêmicos de raça e gênero no ambiente digital.

O estudo permite concluir que o profissional da informação atua como mediador fundamental na zona liminar entre a materialidade técnica e o valor social do dado. Estudos futuros devem expandir esta abordagem para contextos de maior pressão informacional, como a análise de discursos em territórios periféricos, onde as assimetrias e o colonialismo de dados são mais acentuados. Além disso, torna-se necessário aprofundar a investigação a partir de *datasets* mais extensos e processos analíticos de grande escala, de modo a explorar contextos de assimetrias de poder mais acentuados.

REFERÊNCIAS

ARAÚJO, Carlos Alberto Ávila. Uma nova proposta de sistematização da Ciência da Informação: rediscutindo os paradigmas do campo. **Tendências da Pesquisa Brasileira em Ciência da Informação**, [s. *l.*], v. 17, 2024. Disponível em: https://revistas.ancib.org/tpbci/article/view/684. Acesso em: 20 fev. 2026.

BRASIL. Ministério da Ciência, Tecnologia e Inovações. **Estratégia Brasileira para a Transformação Digital 2022-2026**. Brasília, DF: MCTI, 2022. Disponível em: https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/transformacaodigital/arquivosestrategiadigital/e-digital_ciclo_2022-2026.pdf. Acesso em: 3 mar. 2025.

BRASIL. Ministério da Gestão e da Inovação em Serviços Públicos. **Modelo de maturidade de dados**: documento de referência. Versão 2. Brasília, DF: MGI, 2026. Disponível em: https://www.gov.br/governodigital/pt-br/infraestrutura-nacional-de-dados/governancadedados/arquivos/mmd/mmd-documentos/modelo-de-maturidade-de-dados-editorial_publicacao_2026_versao-2.pdf. Acesso em: 15 ago. 2026.

CASELI, Helena de Medeiros; NUNES, Maria das Graças Volpe (org.). **Processamento de linguagem natural**: conceitos, técnicas e aplicações em português. 2. ed. São Carlos: BPLN, 2024.

CASTELLS, Manuel. **A sociedade em rede**. Tradução de Roneide Venancio Majer. 27. ed. Rio de Janeiro: Paz e Terra, 2025.

DAVENPORT, Thomas H.; KIM, Jinho. **Dados demais!**: desmascarando os mitos e descobrindo as oportunidades. Rio de Janeiro: Elsevier, 2014.

GONZÁLEZ DE GÓMEZ, Maria Nélida. Regime de informação: construção de um conceito. **Informação & Sociedade**: Estudos, João Pessoa, v. 22, n. 3, p. 43-60, set./dez. 2012. Disponível em: http://www.ies.ufpb.br/ojs/index.php/ies/article/view/14376. Acesso em: 16 maio 2012.

GONZÁLEZ DE GÓMEZ, Maria Nélida; RABELLO, Rodrigo (org.). **Informação**: agentes e intermediação. Brasília, DF: IBICT, 2017.

HIGGINS, Sarah. Digital curation: the development of a discipline within information science. **Journal of Documentation**, Bingley, v. 74, n. 6, p. 1318-1338, ago. 2018. Disponível em: https://doi.org/10.1108/JD-02-2018-0024. Acesso em: 13 mar. 2024.

INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATIONS AND INSTITUTIONS. **Statement on libraries and artificial intelligence**. [*S. l.*]: IFLA, 2020. Disponível em: https://repository.ifla.org/items/8c05d706-498b-42c2-a93a-3d47f69f7646. Acesso em: 26 fev. 2026.

JOSIOWICZ, Alejandra J.; KOTHARI, Ammina. Digital methods for social media analysis: opportunities and challenges. *In*: JOSIOWICZ, Alejandra J.; VELOZO, Naira de Almeida (org.). **Humanidades digitais na América Latina**: linguagens, metodologias e práticas de análise. 1. ed. Campinas: Pontes Editores, 2023. p. 63-84.

MEMARIAN, Bahar; DOLECK, Tenzin. Fairness, accountability, transparency, and ethics (FATE) in artificial intelligence (AI) and higher education: a systematic review. **Computers and Education**: Artificial Intelligence, Amsterdam, v. 5, 2023. Disponível em: https://www.sciencedirect.com/science/article/pii/S2666920X23000310. Acesso em: 14 jun. 2026.

NOBLE, Safiya Umoja. **Algorithms of oppression**: how search engines reinforce racism. New York: NYU Press, 2018.

PALETTA, Francisco Carlos; SILVA, José Fernando Modesto da; MUCHERONI, Marcos Luiz. Humanidades digitais em ambientes digitais de informação. *In*: OLIVEIRA, Elizângela de Jesus *et al*. (org.). **Tópicos em administração**: volume 41. Belo Horizonte: Poisson, 2021. p. 208-215.

PASQUALE, Frank. **The black box society**: the secret algorithms that control money and information. Cambridge: Harvard University Press, 2015.

PAVÃO, Caterina Groposo *et al*. **Guia de boas práticas para anonimização de dados de pesquisa**. Brasília, DF: IBICT, 2026. *E-book*. Disponível em: https://omp-editora.prd.ibict.br/index.php/edibict/catalog/view/382/551/2893. Acesso em: 10 mar. 2026.

RABELLO, Rodrigo. Documento e institucionalidades: dimensões epistemológica e política. **Encontros Bibli**: revista eletrônica de biblioteconomia e ciência da informação, Florianópolis,

v. 23, n. 51, p. 138-156, 2018. Disponível em: https://doi.org/10.5007/1518-2924.2018v23n51p138. Acesso em: 21 jul. 2026.

RATTENBURY, Tye *et al*. **Principles of data wrangling**: practical techniques for data preparation. Sebastopol: O'Reilly Media, 2017.

ROGERS, David L. **Transformação digital**: repensando o seu negócio para a era digital. Tradução de Afonso Celso da Cunha Serra. São Paulo: Autêntica Business, 2017.

SANT'ANA, Ricardo César Gonçalves. Ciclo de vida dos dados: uma perspectiva a partir da Ciência da Informação. **Informação & Informação**, Londrina, v. 21, n. 2, p. 116-142, 2016. Disponível em: https://ojs.uel.br/revistas/uel/index.php/informacao/article/view/27940. Acesso em: 14 jun. 2025.