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**

# **CAPITU:** ESTRATÉGIAS DE USO NA COLETA DE DADOS NAS MÍDIAS SOCIAIS PARA A PESQUISA ACADÊMICA

## ***CAPITU:*** *STRATEGIES TO COLLECT SOCIAL MEDIA DATA ON ACADEMIC RESEARCH*

**Tainá Batista de Assis** – Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT)
**Josir Cardoso Gomes** – Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT)
**Giulia Tucci** – Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT)
**Bianca da Costa Maia Lope**s – Universidade Federal Fluminense (UFF)
**Graziela Barros Gomes** – Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT*)*

**Modalidade: Trabalho Completo**

**Resumo:** O artigo descreve o desenvolvimento e a consolidação do sistema Capitu, uma infraestrutura de código aberto criada pelo Instituto Brasileiro de Informação em Ciência e Tecnologia desde 2020 para capturar e disponibilizar de forma sistemática dados de mídias sociais para a pesquisa acadêmica. A pesquisa caracteriza-se como um estudo teórico, de natureza qualitativa, com caráter documental, exploratório e descritivo. A relevância da ferramenta insere-se no cenário pós-"APIcalipse", caracterizado pela restrição progressiva e pelo encarecimento do acesso a dados por parte das grandes plataformas digitais, o que impõe sérios obstáculos ao acompanhamento do debate público digital. Construído sob uma abordagem ágil baseada na constante prototipação, a arquitetura da Capitu foi estruturada em quatro eixos: interface para criação de projetos via *queries*, coleta e indexação, processamento e limpeza de metadados e visualização por meio de painéis de gestão (*dashboards*). O estudo conclui destacando a contribuição da Capitu em democratizar a pesquisa independente e promover a integridade da informação sem exigir conhecimentos avançados de programação.
**Palavras-chave:** mídias sociais; recuperação da informação; lago de dados; mecanismos de busca.

**Abstract:** The article describes the development and consolidation of the Capitu system, an open-source infrastructure created by Brazilian Institute of Information in Science and

Technology since 2020 to systematically collect and provide social media data for academic research. The study is theoretical in nature and adopts a qualitative, documentary, exploratory, and descriptive approach. The relevance of the tool is situated within the post-"APIcalypse" context, marked by the progressive restriction and rising costs of access to data imposed by major digital platforms, which pose significant obstacles to monitoring digital public debate. Developed through an agile methodology based on continuous prototyping, Capitu's architecture was structured around four core components: a project creation interface based on queries; data collection and indexing; metadata processing and cleaning; and data visualization through management dashboards. The study concludes by highlighting Capitu's contribution to democratizing independent research and promoting information integrity without requiring advanced programming skills.
**Keywords:** social media; information retrieval; datalakes; search engines.

## 1 INTRODUÇÃO

A crescente digitalização da vida social e a popularização de dispositivos permanentemente conectados transformaram as plataformas digitais em espaços centrais para a circulação de informações, opiniões e disputas simbólicas (Gillespie, 2010). Grandes volumes de dados são produzidos continuamente por usuários, organizações e sistemas automatizados, constituindo uma importante fonte para a compreensão de fenômenos sociais contemporâneos (Boyd; Crawford, 2012).

Paralelamente, observa-se o aumento da circulação de conteúdos enganosos, campanhas coordenadas de manipulação informacional e diferentes formas de desinformação despertam preocupações relacionadas à integridade da informação, à qualidade do debate público e ao funcionamento das instituições democráticas (Benkler; Faris; Roberts, 2018; Bradshaw; Bailey; Howard, 2021).

O estudo desses ambientes digitais depende, entretanto, da disponibilidade de dados produzidos nas plataformas. Durante parte da década de 2010, pesquisadores puderam acessar informações provenientes de redes como X (antigo *Twitter*) e Facebook por meio de interfaces de programação (APIs), conjuntos de regras e protocolos que permitiam a extração automatizada de dados, e outras ferramentas de coleta, o que impulsionou o desenvolvimento de métodos computacionais aplicados às ciências sociais e à Ciência da Informação (CI). Contudo, mudanças nas políticas de governança das plataformas alteraram significativamente esse cenário. Após o escândalo no *Facebook* envolvendo a *Cambridge Analytica* (Confessore, 2018), houve restrição ao acesso a dados por razões associadas à privacidade, segurança e interesses comerciais.

Esse processo se intensificou nos anos seguintes, culminando no "*APIcalipse*" (Bruns, 2019), fenômeno caracterizado pela redução progressiva do acesso a dados de plataformas digitais. O encerramento de programas de acesso acadêmico gratuito no X (antigo *Twitter*), após a aquisição por Elon Musk (Schiffer; Newton; Heath, 2023), é um marco desse movimento. Como consequência, pesquisadores passaram a enfrentar obstáculos crescentes para obter dados necessários à investigação de fenômenos sociais, políticos e informacionais. Tais restrições afetam a capacidade de acompanhamento do debate público digital, de processos como a desinformação e ampliam desigualdades entre instituições e pesquisadores, uma vez que o acesso aos dados passa a depender cada vez mais de recursos técnicos, financeiros e negociações com as próprias plataformas.

Nesse contexto, torna-se fundamental o desenvolvimento de metodologias, ferramentas e infraestruturas de pesquisa capazes de ampliar a acessibilidade à análise de grandes volumes de dados digitais, permitindo que pesquisadores de diferentes áreas explorem e interpretem informações complexas sem a necessidade de conhecimentos avançados em programação ou ciência de dados.

Diante desse quadro de restrições ao acesso a dados de plataformas digitais, torna-se importante que sejam desenvolvidas ferramentas e infraestruturas de pesquisas capazes de ampliar a acessibilidade à análise desses dados digitais, permitindo que pesquisadores de diferentes áreas de conhecimento possam explorar e interpretar informações complexas sem a necessidade de conhecimento e formação avançados em programação ou ciência de dados. Emerge, assim, a seguinte problemática: seria possível construir uma ferramenta que permitisse a pesquisadores da CI e da comunicação coletar dados de redes digitais de forma sistemática, transparente e continuada, sem exigir conhecimentos avançados de programação ou ciência de dados? Essa questão norteadora orienta o presente relato, que tem como objetivo descrever o desenvolvimento do sistema de informação Capitu, que vem sendo desenvolvido pelo Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), desde 2020, com o intuito de capturar e disponibilizar dados de mídias sociais.

Ao descrever o processo de interação entre os pesquisadores que são usuários finais da Capitu e os pesquisadores que o desenvolvem, espera-se que um público mais amplo conheça suas funcionalidades e que mais pesquisadores possam contribuir com o seu aprimoramento.

## 2 REFERENCIAL TEÓRICO

O papel assumido pelas plataformas digitais na mediação das interações sociais e dos fluxos informacionais tem consolidado essas infraestruturas sociotécnicas como objetos centrais de pesquisa para a investigação de uma miríade de fenômenos sociais, políticos e infocomunicacionais contemporâneos (Van Dijck; Poell; De Waal, 2018). Longe de operarem de forma neutra, tais plataformas articulam dispositivos técnicos, lógicas organizacionais e interesses econômicos que influenciam a produção, a visibilidade e a circulação de conteúdos, intervindo ativamente na conformação da esfera pública digital (Gillespie, 2018).

Nesse cenário, a expansão da desinformação, da polarização discursiva, das campanhas coordenadas de manipulação informacional e de outras formas de perturbação estratégica do espaço público (González de Gómez, 2022) impõe novos desafios à investigação científica. Compreender empiricamente as dinâmicas que estruturam esses ambientes exige, por sua vez, procedimentos metodológicos capazes de acompanhar a produção, a circulação e a apropriação da informação em larga escala. A coleta sistemática de dados em redes sociais constitui, nesse sentido, um recurso fundamental para a construção de evidências sobre as relações entre atores, discursos e conteúdo, oferecendo subsídios para a análise de fenômenos que incidem sobre o debate público e sobre o funcionamento das instituições democráticas.

A importância dessas investigações relaciona-se diretamente ao debate contemporâneo sobre integridade da informação. Embora o termo apresente diferentes formulações desde a década de 1980 e ainda se encontre em processo de consolidação conceitual (Yadav; Wanless; Lai, 2026), sua incorporação às agendas de governança de governos e organismos multilaterais têm se intensificado nesta década diante dos desafios decorrentes da crescente complexidade dos ecossistemas informacionais. Nessa acepção mais recente, a integridade da informação associa-se à promoção de ambientes caracterizados pela precisão, consistência e confiabilidade das informações que circulam na esfera pública (ONU, 2023).

Sob esse prisma, a integridade da informação extrapola a tradição da segurança da informação – em sua concepção conceitual originalmente vinculada a atributos meramente técnicos – e passa a incorporar dimensões sociais, políticas e epistemológicas relacionadas à confiabilidade dos fluxos informacionais e às condições de acesso a informações verificáveis e

contextualizadas (Lopes, 2025). Trata-se, portanto, de compreendê-la como uma condição resultante de processos sociotécnicos que favorecem a produção, a organização, a circulação e o uso social da informação. Em oposição à corrupção informacional, seu conceito aponta para a capacidade de sustentar ecossistemas informacionais transparentes e confiáveis, condição fundamental para o fortalecimento da resiliência social (OCDE, 2024) e da qualidade da vida democrática.

A pesquisa acadêmica assume posição estratégica nesse debate ao produzir evidências científicas sobre processos que comprometem a integridade dos ambientes informacionais e, em última instância, da própria democracia. Contudo, para que tais análises obtenham consistência metodológica é indispensável dispor de instrumentos que assegurem rastreabilidade, transparência e possibilidade de verificação dos procedimentos de coleta e tratamento dos dados. Sobretudo, importa que esses instrumentos sejam acessíveis e inteligíveis para os pesquisadores.

Essa preocupação encontra respaldo nos Princípios Globais para a Integridade da Informação das Nações Unidas (ONU, 2024), que reconhecem a transparência e a pesquisa independente como pilares dos ecossistemas informacionais. Ao defender a ampliação do acesso da comunidade científica a dados sobre o funcionamento das plataformas digitais, o documento reconhece que a promoção da integridade da informação depende igualmente da atuação da academia, da sociedade civil e das instituições públicas na produção de conhecimento e no escrutínio dos ambientes digitais.

Nessa perspectiva, as ferramentas de coleta de dados em redes sociais não se restringem a instrumentos técnicos de pesquisa. Ao viabilizarem maior transparência, rastreabilidade e reprodutibilidade dos procedimentos investigativos, tornam-se recursos fundamentais para a produção de evidências sobre os ecossistemas informacionais digitais. Contribuem, assim, tanto para o avanço do conhecimento científico quanto para o fortalecimento da integridade da informação como bem público, ampliando a capacidade analítica da academia diante dos desafios informacionais que afetam a sociedade contemporânea.

A materialização dessas possibilidades investigativas depende, todavia, das condições de acesso aos dados disponibilizados pelas plataformas digitais; em particular, da existência de mecanismos que permitam acessar, coletar e organizar os dados produzidos nesses ambientes. Os próprios dispositivos, objetos e mecanismos nativos da web podem ser usados

para investigar fenômenos sociais e culturais (Rogers, 2013). Historicamente, um dos principais mecanismos de acesso a esses dados foi por meio das APIs. APIs consistem em conjuntos de regras e protocolos que permitem a comunicação entre sistemas computacionais, possibilitando que pesquisadores consultem conteúdos, metadados e informações sobre interações de forma automatizada. Durante muitos anos, plataformas como X (antigo *Twitter*) e *Facebook* disponibilizaram APIs que permitiam a coleta sistemática de dados para fins acadêmicos (Venturini; Rogers, 2019), favorecendo o desenvolvimento de pesquisas em áreas como comunicação, ciência da informação, sociologia e ciência política.

Entretanto, as APIs não constituem mecanismos homogêneos. Cada plataforma estabelece diferentes níveis de acesso, limites de requisição, formatos de dados e critérios de disponibilização de informações. Como consequência, os objetos de pesquisa acessíveis e as análises possíveis variam significativamente entre plataformas. Enquanto algumas permitem a recuperação de conteúdos, comentários, métricas de engajamento e redes de interação, outras restringem o acesso a determinadas informações ou impõem barreiras geográficas, técnicas e financeiras para sua utilização.

Além das APIs oficiais, pesquisadores frequentemente recorrem a bibliotecas e ferramentas desenvolvidas por terceiros para acessar dados disponibilizados pelas plataformas. No caso do Telegram, por exemplo, a biblioteca *Telethon* ([*s.d.*]) permite a coleta de mensagens, mídias e metadados de grupos e canais públicos a partir da API oficial da plataforma. O *YouTube*, por sua vez, oferece acesso a informações públicas por meio da API oficial, e o *YouTube Data Tools* ([s.d.]) possibilita a coleta de *links* para vídeos, comentários, canais e métricas associadas. Em contrapartida, outras plataformas apresentam cenários mais restritivos. O acesso a dados do X depende atualmente de modalidades pagas de utilização da API, enquanto plataformas como TikTok oferecem possibilidades limitadas de acesso para pesquisa acadêmica, especialmente fora dos programas institucionais específicos disponibilizados pela empresa. Já a corporação Meta, detentora das plataformas Facebook e Instagram, oferece a ferramenta *Meta Content Library* para pesquisadores que se cadastram na plataforma. Mesmo após várias tentativas, infelizmente os pesquisadores da Capitu nunca conseguiram acesso a esse serviço.

Ao longo da pesquisa, foram encontradas várias ferramentas proprietárias que coletam dados de redes sociais sem a necessidade de programação, entretanto o custo de aquisição é extremamente alto para uso na pesquisa acadêmica. Além disso, tais serviços não

informam como a coleta é realizada ou se os dados que estão sendo coletados serão utilizados por terceiros. Desta forma, investir em uma ferramenta de código aberto e transparente pode abrir inúmeras possibilidades para pesquisadores que necessitam coletar dados das mídias sociais.

## 3 PROCEDIMENTOS METODOLÓGICOS

A presente pesquisa caracteriza-se como um estudo teórico, de natureza qualitativa, com caráter documental, exploratório e descritivo. A abordagem qualitativa fundamenta a análise do desenvolvimento do sistema de informação Capitu à luz do arcabouço teórico da Ciência da Informação, especialmente no que se refere à coleta de dados provenientes de mídias sociais e às diversas formas de apoiar pesquisas acadêmicas relacionadas ao tema.

A presente pesquisa caracteriza-se como um estudo teórico, de natureza qualitativa, com caráter documental, exploratório e descritivo. Seu caráter documental decorre da análise de documentação técnica, registros de desenvolvimento, painéis de gestão e demais artefatos produzidos ao longo da trajetória da Capitu, utilizados como fontes primárias para a reconstrução de seu processo de desenvolvimento. O caráter exploratório justifica-se pela escassez de sistematizações prévias sobre infraestruturas de coleta de dados de mídias sociais desenvolvidas no contexto pós-APIcalipse, sobretudo as de código aberto e voltadas à realidade da pesquisa acadêmica brasileira, o que exige uma primeira aproximação descritiva do objeto antes de análises comparativas ou avaliativas mais amplas. Já o caráter descritivo relaciona-se ao objetivo de detalhar, de forma sistemática e estruturada, a arquitetura, os componentes e o funcionamento da ferramenta, sem a pretensão de testar hipóteses ou estabelecer relações causais entre variáveis (Gil, 2017). Embora estudos exploratórios e descritivos frequentemente se sobreponham, à medida que toda descrição inicial de um fenômeno pouco sistematizado tem também função exploratória, optou-se por manter ambas as classificações para indicar tanto a natureza ainda incipiente da literatura sobre o objeto quanto o compromisso com uma caracterização estruturada e replicável de seus componentes.

Para a descrição do sistema, adotou-se como estratégia metodológica a decomposição da Capitu em eixos funcionais, de modo a estruturar a apresentação dos resultados e permitir sua compreensão sistemática por parte de pesquisadores não familiarizados com sua arquitetura. Foram definidos quatro eixos de análise, construídos a partir da própria lógica de

desenvolvimento do sistema e validados pela equipe técnica responsável por sua implementação: i) interface de criação de projetos e definição de *queries*; ii) coleta e indexação de dados; iii) processamento e limpeza de metadados; e iv) visualização por meio de painéis de gestão. A descrição de cada eixo baseou-se em observação participante da equipe de desenvolvimento, análise da documentação técnica do sistema (incluindo o repositório de código aberto) e registros quantitativos de uso extraídos dos próprios *datalakes* da ferramenta. Essa triangulação de fontes permitiu descrever os componentes técnicos da Capitu, bem como situar sua trajetória de desenvolvimento em relação aos desafios de escalabilidade e sustentabilidade enfrentados ao longo dos seis anos de operação.

A coleta de dados em ambiente digital é, por natureza, uma atividade interdisciplinar que combina práticas de inteligência de fontes abertas (OSINT) e técnicas de recuperação da informação. De acordo com Glassman e Kang (2012), a OSINT emerge como abordagem sistemática para a produção de conhecimento a partir de dados publicamente disponíveis na internet, integrando coleta, processamento, análise e disseminação.

Para a construção do sistema de informação, utilizou-se a metodologia PAR (Rocha; Alves; Braga, 2023) que é baseada em metodologias ágeis onde o sistema é desenvolvido a partir de constante prototipação e interação com os usuários. Esta metodologia tem sido utilizada com sucesso em diversos sistemas desenvolvidos nos IBICT.

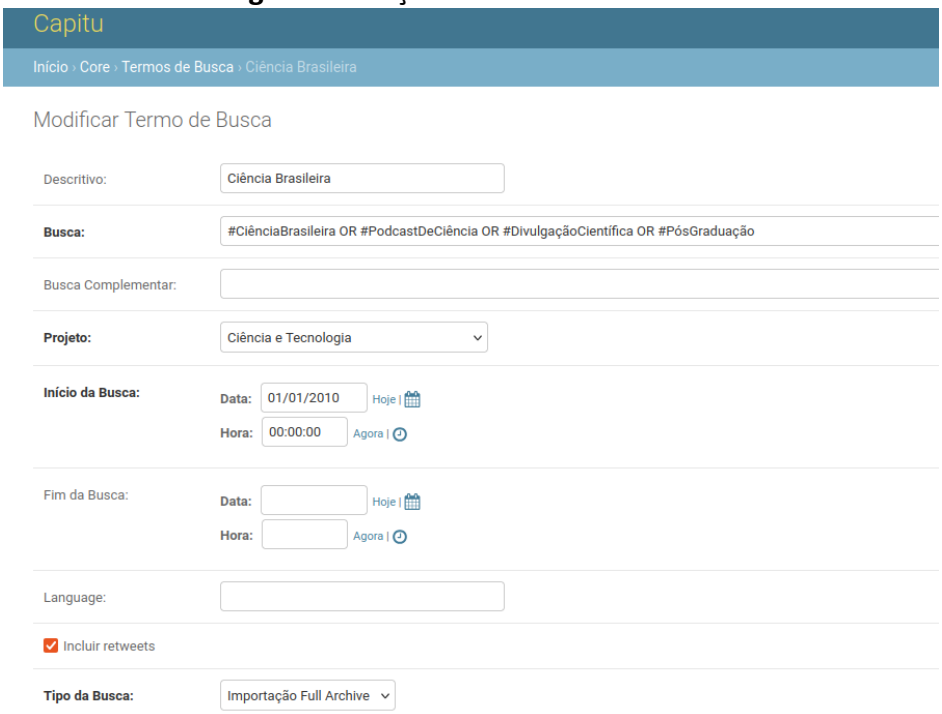
## 4 RESULTADOS

A apresentação dos resultados está organizada segundo os quatro eixos que estruturam a arquitetura da Capitu, tal como anunciado na metodologia, detalhados a seguir: (i) a interface para a criação dos projetos onde são definidos os parâmetros de coleta, em especial as queries de busca, das listas de canais e de quais redes serão acessadas; (ii) a coleta dos dados das plataformas digitais através das APIs e o armazenamento e a indexação dos dados coletados; (iii) o processamento dos dados coletados, incluindo o mapeamento de metadados e a aplicação de filtros que não estão disponíveis nas APIs; (iv) a visualização, o controle de acesso e a interação dos pesquisadores com os dados coletados, mediada por dashboards, painéis web e pela própria interface da Capitu.

## 4.1 INTERFACE PARA CRIAÇÃO DE PROJETOS E DEFINIÇÃO DE *QUERIES*

Adota-se, no contexto da Capitu, a definição operacional segundo a qual uma query corresponde ao escopo que compõe as sentenças de busca e os parâmetros de busca utilizados para a coleta de dados em diferentes plataformas digitais. Esta definição é fundamental porque a query é o principal instrumento de delimitação do objeto empírico e, portanto, condiciona a validade interna dos achados (Howison; Wiggins; Crowston, 2011).

**Figura 1** - Edição dos termos de busca




**Fonte**: dados da pesquisa (2026).

Na Figura 1, pode-se visualizar uma das telas onde os usuários podem registrar as *queries*.

## 4.2 COLETA, ARMAZENAMENTO E INDEXAÇÃO DOS DADOS

Uma vez que as *queries* são definidas, o sistema dispara as funções de captura de dados utilizando as APIs de cada uma das plataformas e os dados coletados são armazenados em formato bruto em *datalakes* (Hai *et al.*, 2023) para cada plataforma, ou seja, em uma estrutura de dados específica para o *Twitter*, um para o *Telegram* e assim sucessivamente. A opção de se utilizar *datalake* segregados se deu em função do formato de dados que cada API retorna e também para garantir uma maior velocidade para armazená-los.

O armazenamento é realizado nos bancos de dados *MariaDB*, *PostgreSQL* e *OpenSearch*. Inicialmente, utilizava-se apenas o banco *MariaDB* em um formato relacional, mas com o aumento expressivo do volume de dados, esta arquitetura não se mostrou rápida o suficiente para as consultas que eram realizadas pelos usuários. Dessa forma, optou-se por migrar os dados para o *Opensearch*. A migração foi feita em função de sua capacidade de busca de alta performance, consultas estruturadas, de permitir armazenar grandes volumes de dados e do modelo de licenciamento, baseado em licenças livres.

Ao longo destes seis anos, somente no X/*Twitter* foram coletadas 8,5 milhões de postagens e 15 milhões de *retweets* e comentários. Também foram coletados dados públicos de mais de 4,8 milhões de usuários em mais de 70 projetos de pesquisa. Já a coleta do *YouTube* foi mais modesta pois a coleta começou apenas em 2025 e não houve tanta procura por parte dos pesquisadores: foram coletadas estatísticas de aproximadamente 342 mil vídeos.

Até a metade de 2025, o volume de dados coletados girava em torno de 50 mil registros por mês. Entretanto, ao longo do ano do uso da Capitu em projetos de pesquisa do IBICT, este número saltou para quase 1 milhão de registros por mês. Neste ponto, a infraestrutura computacional precisou ser redimensionada já que o *Opensearch* não suportou armazenar tanto os *datalakes* de cada rede como os índices de busca.

Por conta desta reestruturação, somente os módulos de coleta das plataformas X/*Twitter* e *YouTube* estão habilitados. As capturas de *Telegram* e *Bluesky* foram interrompidas para que se consiga otimizar as cargas nos *datalakes*. Ainda sobre os bancos de dados, o uso de metadados padronizados facilita integrações e análises posteriores, além de permitir que diferentes produtos informacionais acessem o mesmo repositório comum de forma eficiente.

Especificamente para o *Telegram*, a estratégia de coleta é diferenciada já que ele não permite a busca por *queries*, tal como nas outras três plataformas. Dessa forma, o sistema coleta dados de canais públicos selecionados, gravando no *datalake* específico do *Telegram*. A partir dos dados coletados, são realizadas as *queries* no *datalake* para em seguida disponibilizar os dados para cada projeto.

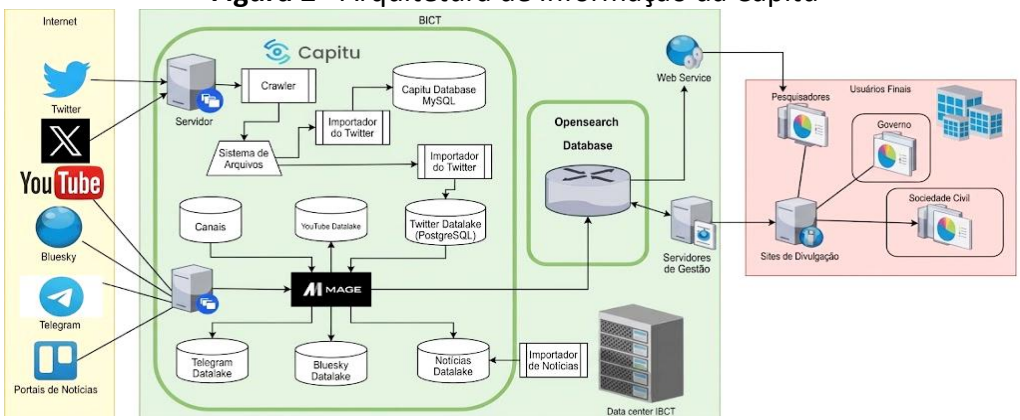
Sob a perspectiva da CI, a definição de *queries* na Capitu trata-se de um processo de representação e delimitação do universo informacional, no qual o pesquisador negocia categorias temáticas, operadores lógicos e recortes temporais que condicionam a temática a ser analisada. Esse movimento dialoga diretamente com a tradição dos estudos de

recuperação da informação, com a noção de estratégia de busca como processo iterativo de construção do conhecimento (Kuhlthau, 2013) e com a compreensão de que os dispositivos nativos da *web* constituem objetos legítimos de investigação social e cultural (Rogers, 2013). dispensar conhecimentos avançados de programação, a Capitu amplia o acesso de pesquisadores da CI à coleta sistemática de dados de mídias sociais, em um contexto de restrição crescente de APIs.

## 4.3 PROCESSAMENTO E LIMPEZA DE METADADOS

O uso de metadados padronizados facilita integrações e análises posteriores, além de permitir que diferentes produtos informacionais acessem o mesmo repositório comum de forma eficiente, contribuindo para a consistência dos dados disponibilizados aos pesquisadores.

**Figura 2** - Arquitetura de Informação da Capitu




**Fonte**: dados da pesquisa (2026).

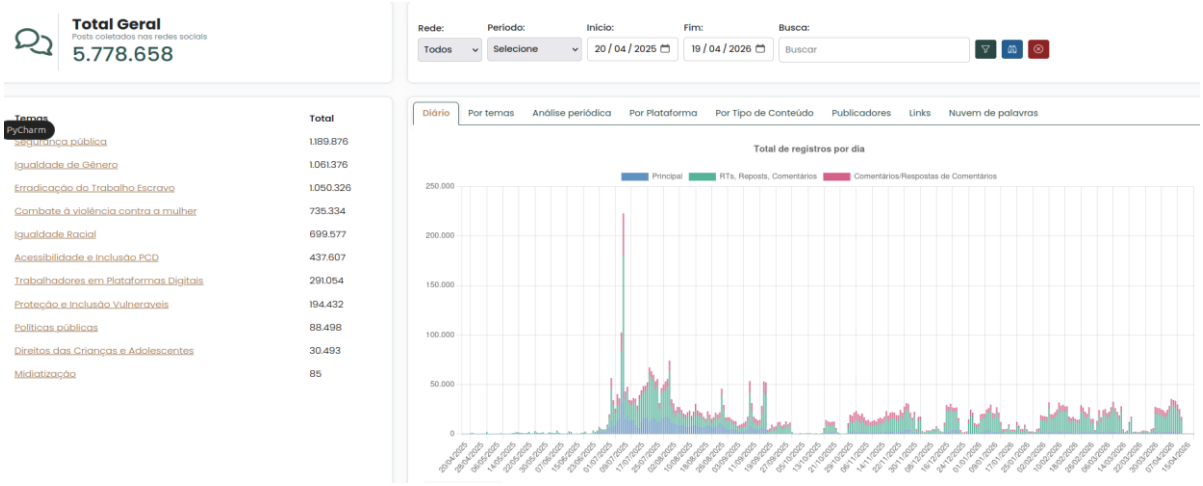
A Figura 2 apresenta a arquitetura de informação completa da Capitu, evidenciando a relação entre os módulos de coleta, os *datalakes* segregados por plataforma, o processamento de metadados e o banco OpenSearch.

## 4.4 VISUALIZAÇÃO, CONTROLE DE ACESSO E INTERAÇÃO DOS PESQUISADORES

Por fim, os dados são também disponibilizados em um ambiente de gestão de dados, no qual pesquisadores podem ter uma visão macro dos dados coletados por período, classificar publicações, anotar evidências e refinar continuamente os termos de busca. Esse ciclo de governança é constitutivo da estratégia metodológica: as *queries* são revisadas em função das dinâmicas observadas no debate público digital, em um movimento iterativo de

calibração entre o escopo temático e materialidade discursiva. Na Figura 3, é possível visualizar um dos painéis de consulta.

**Figura 3** - Painel de gestão




**Fonte**: dados da pesquisa (2026).

O módulo de coleta já tem o código aberto e está disponível no github[1]. Já os novos módulos de carga do *datalake* e o ambiente de gestão também estarão abertos assim que estiverem finalizados. A exportação de resultados já se encontra implementada em ambiente interno da Capitu e está prevista para disponibilização futura aos pesquisadores usuários.

A arquitetura da Capitu pode ser compreendida como uma infraestrutura informacional que articula diferentes problemas tradicionalmente investigados pela Ciência da Informação. A definição de *queries* envolve processos de busca, recuperação e delimitação do universo informacional; a coleta e o armazenamento dos dados remetem à organização e à gestão de grandes volumes de informação; o tratamento e a padronização de metadados favorecem consistência, interoperabilidade e reutilização; e os painéis de visualização e gestão operam como dispositivos de mediação da informação, ao traduzirem grandes volumes de dados em representações visuais que favorecem a interpretação, a comparação e a identificação de padrões.

Essa função aproxima a ferramenta das discussões sobre apropriação informacional e sobre as condições de acesso à informação verificável e contextualizada (González de Gómez, 2022; Lopes, 2025), ao permitir que o pesquisador classifique publicações, anote evidências e refine os termos de busca. Ao integrar essas etapas em uma infraestrutura aberta e documentada, a Capitu amplia a rastreabilidade e a transparência dos procedimentos de

[1] Disponível em: https://github.com/josircg/twitsearch.

pesquisa com dados de plataformas digitais, condições fundamentais para a integridade da informação como bem público (ONU, 2024; OCDE, 2024).

## 5 CONSIDERAÇÕES FINAIS

A Capitu foi desenvolvido inicialmente pelo Laboratório em Rede de Humanidades Digitais (LARHUD) e desde então tem sido utilizado em diversos projetos acadêmicos (Gomes; Pimenta, 2022) e ao longo de seu desenvolvimento e da interação com os pesquisadores que o utilizaram, foi possível avaliar as inúmeras possibilidades de uso da Capitu na pesquisa acadêmica e a existência de inúmeras novas funcionalidades que podem ser incorporadas. Além disso, o painel de gestão foi consolidado e recebeu inúmeras melhorias a partir das interações com os pesquisadores que participaram das coletas.

Os resultados apresentados indicam que o desenvolvimento da Capitu responde a uma demanda técnica por acesso a dados de plataformas digitais e a questões informacionais relacionadas à recuperação, organização, tratamento, acesso e uso de dados digitais para pesquisa. Nesse sentido, sua principal contribuição para a Ciência da Informação está na proposição e consolidação de uma infraestrutura que torna mais transparentes e sistemáticos diferentes estágios da construção de dados de pesquisa, particularmente em um contexto marcado pela crescente restrição do acesso às plataformas digitais.

Sob o ponto de vista da coleta, o acesso a dados do X/Twitter ainda restringe muito seu uso devido ao alto custo da API de modo que as estratégias de criação das consultas (*queries*) devem ser construídas com muita atenção. Além disso, tem havido uma conscientização maior dos gestores dos projetos de pesquisa para que eles fiquem atentos a incluir o custo destas APIs nos orçamentos de suas pesquisas.

Busca-se agora novas fontes de financiamento para que se consiga construir mecanismos de coleta de portais de notícias e para as novas mídias sociais abertas como o *Mastodon* e a *PixelFed*, ambas encontradas no Fediverso (Gomes, 2023). Com coletas mais amplas, espera-se que os pesquisadores tenham mais possibilidades na escolha de seus métodos.

Entre as limitações, destacam-se a dependência das políticas e condições técnicas impostas pelas próprias plataformas, os custos associados a determinados mecanismos de acesso, particularmente no caso do X, e a disponibilidade ainda desigual dos módulos de coleta entre diferentes plataformas digitais. Embora o desenvolvimento da Capitu tenha

ocorrido de maneira iterativa e em interação com pesquisadores usuários da ferramenta, ainda não foi realizada uma avaliação sistemática de usabilidade e percepção desses usuários. Estudos futuros poderão incorporar esse tipo de avaliação, examinar comparativamente a qualidade e a cobertura dos dados obtidos e acompanhar a expansão da infraestrutura para novas fontes e plataformas.

## REFERÊNCIAS

BENKLER, Yochai; FARIS, Robert; ROBERTS, Hal. Network propaganda: manipulation, disinformation, and radicalization in American politics. New York: **Oxford University Press**, 2018. Disponível em: https://academic.oup.com/book/26406. Acesso em: 12 set. 2026.

BOYD, Danah; CRAWFORD, Kate. Critical questions for Big Data: provocations for a cultural, technological, and scholarly phenomenon. **Information, Communication & Society**, v. 15, n. 5, p. 662–679, 2012. Disponível em: https://www.tandfonline.com/doi/full/10.1080/1369118X.2012.678878. Acesso em: 8 set. 2026.

BRADSHAW, Samantha; BAILEY, Hannah; HOWARD, Philip N. Industrialized disinformation: 2020 global inventory of organized social media manipulation. Working Paper 2021.1. Oxford: **Project on Computational Propaganda**, 2021. Disponível em: https://www.oii.ox.ac.uk/news-events/reports/industrialized-disinformation-2020-global-inventory-of-organized-social-media-manipulation/. Acesso em: 11 nov. 2026.

BRAGA, Tiago Emmanuel Nunes; BRANDÃO, Renam Vinicius Carvalho Santos; ASSIS, Tainá Batista de; ALVES, Larissa de Araújo; GOMES, Graziela Barros; SILVA, Rodrigo Marcelino da; CALIARI, Rafael Marques. Difusão: primeiros passos de uma proposta informacional para o combate à desinformação. *In*: **WORKSHOP DE INFORMAÇÃO, DADOS E TECNOLOGIA**, 7., 2024, Porto Velho. Anais [...]. [S. l.]: IBICT, 2024. Disponível em: https://widat.ibict.br/widat2024/article/view/175. Acesso em: 8 set. 2026.

BRUNS, Axel. After the 'APIcalypse': social media platforms and their fight against critical scholarly research. **Information, Communication & Society**, v. 22, n. 11, p. 1544–1566, 2019. Disponível em: https://www.tandfonline.com/doi/abs/10.1080/1369118X.2019.1637447. Acesso em: 8 set. 2026.

CONFESSORE, Nicholas. Cambridge Analytica and Facebook: the scandal and the fallout so far. **The New York Times**, 4 abr. 2018. Disponível em: https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html. Acesso em: 8 set. 2026.

GIL, Antonio Carlos. **Como elaborar projetos de pesquisa**. 6. ed. São Paulo: Atlas, 2017. Disponível em: https://biblioteca.iftm.edu.br/acervo/detalhe/20572. Acesso em: 11 nov. 2026.

GILLESPIE, Tarleton. Custodians of the Internet: platforms, content moderation, and the hidden decisions that shape social media. New Haven: **Yale University Press**, 2018. Disponível em: https://yalebooks.yale.edu/book/9780300235029/custodians-of-the-internet/. Acesso em: 11 nov. 2026.

GILLESPIE, Tarleton. The politics of 'platforms'. **New Media & Society**, v. 12, n. 3, p. 347–364, 2010. Disponível em: https://journals.sagepub.com/doi/10.1177/1461444809342738. Acesso em: 8 set. 2026.

GLASSMAN, Michael; KANG, Min Ju. Intelligence in the internet age: The emergence and evolution of Open Source Intelligence (OSINT). **Computers in Human Behavior**, v. 28, n. 2, p. 673-682, mar. 2012. Disponível em: https://doi.gov10.1016/j.chb.2011.11.014. Acesso em: 13 nov. 2026.

GOMES, Josir Cardoso. Ressignificando as redes sociais: a rede Mastodon sob a ótica da economia política da informação. *In*: **ENCONTRO NACIONAL DE PESQUISA E PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO, 23**, 2023, Aracaju. Anais [...]. [S. l.]: ANCIB, 2023. Disponível em: https://www.enancib.ancib.org/enancib/en/article/view/486. Acesso em: 5 set. 2024.

GOMES, Josir Cardoso; PIMENTA, Ricardo Medeiros. Impactos da ferramenta Twitter Monitor para pesquisas acadêmicas sobre redes sociais digitais. In: CONGRESSO ALAIC, 16., 2022, Buenos Aires. Anais [...]. Buenos Aires: **ALAIC**, 2022. Disponível em: https://doi.org/10.5281/zenodo.7458731. Acesso em: 8 set. 2026.

GONZÁLEZ DE GÓMEZ, Maria Nélida. Orientações contemporâneas da Ciência da Informação: vinculações com a epistemologia social. **Museologia & Interdisciplinaridade**, Brasília, v. 11, n. 22, p. 179–198, 2022. Disponível em: https://periodicos.unb.br/index.php/museologia/article/view/43380. Acesso em: 8 set. 2026.

HAI, Rihan; KOUTRAS, Christos; QUIX, Christoph; JARKE, Matthias. Data lakes: a survey of functions and systems. **IEEE Transactions on Knowledge and Data Engineering**, v. 35, n. 12, p. 12571–12590, 2023. Disponível em: https://ieeexplore.ieee.org/document/10107808/. Acesso em: 8 set. 2026.

HOWISON, James; WIGGINS, Andrea; CROWSTON, Kevin. Validity issues in the use of social network analysis with digital trace data. **Journal of the Association for Information Systems,** v. 12, n. 12, p. 767–797, 2011. Disponível em: https://aisel.aisnet.org/jais/vol12/iss12/2/. Acesso em: 8 set. 2026.

KUHLTHAU, Carol. **Como usar a biblioteca na escola**: um programa de atividades para o ensino fundamental. 3. ed. Belo Horizonte: Autêntica, 2013. Disponível em: http://grupoautentica.com.br/autentica/livros/como-usar-a-biblioteca-na-escola-um-programa-de-atividades-para-o-ensino-fundamental/51. Acesso em: 11 nov. 2026.

LOPES, Bianca da Costa Maia. Integridade da informação: construção crítica do conceito na

contemporaneidade. Encontros Bibli: **revista eletrônica de biblioteconomia e ciência da informação**, Florianópolis, v. 31, e106937, 2026. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/106937. Acesso em: 8 set. 2026.

ORGANIZAÇÃO PARA A COOPERAÇÃO E DESENVOLVIMENTO ECONÔMICO (OCDE). Recommendation of the Council on Information Integrity. Paris: **OECD**, 2024. Disponível em: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0505. Acesso em: 15 jun. 2026.

ORGANIZAÇÃO DAS NAÇÕES UNIDAS (ONU). Our Common Agenda: policy brief 8: information integrity on digital platforms. New York: **United Nations**, 2023. Disponível em: https://digitallibrary.un.org/record/4012857. Acesso em: 15 jun. 2026.

ORGANIZAÇÃO DAS NAÇÕES UNIDAS (ONU). United Nations Global Principles for Information Integrity: recommendations for multi-stakeholder action. New York: **United Nations**, 2024. Disponível em: https://digitallibrary.un.org/record/4052524. Acesso em: 15 jun. 2026.

ROCHA, Nathaly; ALVES, Larissa de Araújo; BRAGA, Tiago Emmanuel Nunes. Metodologia colaborativa para levantamento de necessidades informacionais em projetos de pesquisa. *In*: **WORKSHOP DE INFORMAÇÃO, DADOS E TECNOLOGIA**, 6., 2023, Brasília. Anais [...]. [*S. l.*]: IBICT, 2023. p. 1–13. Disponível em: https://labcotec.ibict.br/widat/index.php/widat2023/article/view/76. Acesso em: 22 jun. 2026.

ROGERS, Richard. **Digital methods**. Cambridge, MA: MIT Press, 2013. Disponível em: https://mitpress.mit.edu/9780262313391/digital-methods/. Acesso em: 11 nov. 2026.

SCHIFFER, Zoë; NEWTON, Casey; HEATH, Alex. Tears, blunders and chaos: inside Elon Musk's Twitter. **The Guardian**, 29 jan. 2023. Disponível em: https://www.theguardian.com/technology/2023/jan/29/tears-blunders-and-chaos-inside-elon-musk-twitter. Acesso em: 22 jun. 2026.

TELETHON. Telethon 1.30.3 documentation. [S. l.]: **Telethon**, [s. d.]. Disponível em: https://docs.telethon.dev/en/stable/. Acesso em: 3 jan. 2024.

VAN DIJCK, José; POELL, Thomas; DE WAAL, Martijn. The Platform Society: public values in a connective world. **New York: Oxford University Press**, 2018. Disponível em: https://academic.oup.com/book/12378. Acesso em: 11 nov. 2026.

VENTURINI, Tommaso; ROGERS, Richard. "API-based research" or how can digital sociology and journalism studies learn from the Facebook and Cambridge Analytica data breach. **Digital Journalism**, v. 7, n. 4, p. 532–540, 2019. Disponível em: https://www.tandfonline.com/doi/abs/10.1080/21670811.2019.1591927. Acesso em: 20 jun. 2026.

YOUTUBE DATA TOOLS. YouTube Data Tools. [S. l.]: **Digital Methods Initiative**, [*S. d.*].

Disponível em: https://ytdt.digitalmethods.net/. Acesso em: 5 abr. 2024.

YADAV, Kamya; WANLESS, Alicia; LAI, Samantha. Multilateral efforts on information integrity: why a clear definition is needed. In: GINSBORG, Lisa; GORI, Paula (ed.). Disinformation: a multi-disciplinary analysis. Cham: **Springer**, 2026. p. 489–511. Disponível em: https://link.springer.com/chapter/10.1007/978-3-032-00480-2_25. Acesso em: 15 jun. 2026.

**FINANCIAMENTO**

O aprimoramento da Capitu está sendo financiado pelo Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict) por meio do projeto intitulado “Infraestrutura pública aberta para o fortalecimento da integridade da informação a partir da coleta dos dados das plataformas digitais”.