

Uma reflexão sobre história digital e soberania: o uso e reuso de bancos de dados

A reflection on digital history and sovereignty: the use and reuse of databases

Alesson Ramon Rota

Doutor em História pela Universidade Estadual de Campinas, Brasil
Pós-doutorando na Universidade Estadual Paulista, Brasil
alesson.rota@gmail.com
<https://orcid.org/0000-0001-9167-7903>
<http://lattes.cnpq.br/3896360099405972>

Resumo: O artigo investiga como produzir história digital frente ao colonialismo de dados, entendido como a articulação entre extração, circulação desigual, classificação e governança assimétrica de dados e infraestruturas digitais. Para responder a essa questão, o texto organiza três eixos: o significado da história praticada hoje, quando a evidência passa por bancos de dados, metadados, plataformas e código; o significado do digital, não somente como suporte, mas como meio técnico, econômico e político que condiciona a visibilidade dos rastros; e as estratégias de autonomia relativa possíveis para países latino-americanos. Situando o conceito de soberania e dialogando com críticas à datificação, sustenta que recortes de dados e métodos computacionais produzem efeitos epistêmicos e políticos, exigindo crítica de fontes, transparência e governança pública de infraestrutura. A argumentação se apoia no reuso de cinco conjuntos de dados/fontes digitais: contratos públicos brasileiros com Big Techs; mapeamentos da concentração global de capacidade de IA; indicadores sobre extrativismo e economia digital; metadados SciELO, Crossref e OpenAlex para georreferenciar produção latino-americana; e um dataset de publicações da Plataforma X arquivado em repositório acadêmico. Ao final, defende ciência aberta, reuso de dados e soluções de código aberto como práticas capazes de criar defesas parciais contra dependências proprietárias e colonialismos digitais.

Palavras-chaves: soberania; história digital; algoritmos; ciência aberta; infraestrutura.

Abstract: The article investigates how to produce digital history in the face of data colonialism, understood as the articulation between extraction, unequal circulation, classification, and the asymmetric governance of data and digital infrastructures. To address this question, the text is organized around three axes: the meaning of history as practiced today, when evidence passes through databases, metadata, platforms, and code; the meaning of the digital, not merely as a medium of storage, but as a technical, economic, and political environment that conditions the visibility of traces; and the strategies of relative autonomy

available to Latin American countries. By situating the concept of sovereignty and engaging with critiques of datafication, the article argues that data selections and computational methods produce epistemic and political effects, requiring source criticism, transparency, and the public governance of infrastructure. The argument is based on the reuse of five datasets/digital sources: Brazilian public contracts with Big Tech companies; mappings of the global concentration of AI capacity; indicators on extractivism and the digital

economy; SciELO, Crossref, and OpenAlex metadata used to georeference Latin American scholarly production; and a dataset of posts from Platform X archived in an academic repository. In conclusion, the article advocates open science, data reuse, and open-source solutions as practices capable of creating partial defenses against proprietary dependencies and digital colonialisms.

Keywords: sovereignty; digital history; algorithms; open science; infrastructure.

Introdução

O presente artigo tem como objetivo responder a uma pergunta central: como produzir história digital frente ao chamado colonialismo de dados ou colonialismo digital? A questão não se limita ao uso de ferramentas computacionais para organizar documentos; ela exige compreender como dados, metadados, plataformas, contratos de nuvem, linguagens de programação, infraestruturas de processamento e regimes de preservação participam da produção da evidência histórica.

Embora seja uma questão importante para as discussões sobre soberania digital, também cara a produção de uma história digital que explora as potencialidades computacionais, os seus desafios não obstam a produção de história, sobretudo se entendemos em perspectiva historiográfica as noções envolvidas. História digital, colonialismo de dados, soberania digital, sinalizam o estreitamento em relação a computação, mas se removemos o verniz da computação moderna, pendêramo-nos com desafios já conhecidos.

Por isso, o argumento se organiza em três movimentos. Primeiro, problematiza demandas da história praticada quando a crítica documental precisa alcançar bases de dados, código, plataformas e cadeias de custódia digitais. Segundo, pergunta o que significa o digital: um meio técnico que condiciona escalas de observação, formas de acesso, padrões de visibilidade e dependências materiais. Terceiro, examina quais estratégias podem criar autonomia relativa para países latino-americanos, especialmente por meio de ciência aberta, reuso de dados, software livre, código aberto e infraestruturas públicas de preservação. A América Latina aparece, portanto, como localização analítica do problema e como lugar de produção de alternativas.

Ao passo que o conceito de soberania se revela histórico, ele reaparece no vocabulário do debate público a partir da popularização da Inteligência Artificial (IA). A produção de sistemas de IA implica as dinâmicas entre os Estados nacionais para aquisição dos diferentes recursos para seu treinamento e inferência (uso). Se antes era apenas uma aplicação científica de elevados custos ou matéria de ficção científica, a IA ganhou diferentes contornos a partir do lançamento do ChatGPT em 30 de novembro de 2022.¹ Para a história digital, interessa como a política internacional implica o uso computacional na América Latina, tema caro para o desenvolvimento deste emergente campo da escrita da história.

Neste artigo, a “soberania digital” é discutida como um conceito operacional para avaliar as condições materiais e institucionais da pesquisa histórica quando mediada por infraestruturas digitais. Ela diz respeito à capacidade de atores públicos (Estado, universidades e consórcios) de governar, de modo transparente e auditável, a infraestrutura de armazenamento, processamento e preservação, os regimes de acesso e custo (contratos e serviços em nuvem) e as rotinas metodológicas que produzem evidência a partir de bases de dados (documentação, metadados e reprodutibilidade).²

A partir desse enquadramento, o texto sustenta três hipóteses: (1) soberania digital é uma condição infraestrutural e de governança, e não apenas um princípio abstrato; (2) bancos de dados e algoritmos não apenas “representam” o passado, mas participam da fabricação e estabilização da evidência, exigindo uma crítica de fontes estendida a metadados, plataformas e opacidades técnicas; e (3) a dependência computacional reatualiza assimetrias do capitalismo global (energia, mineração, nuvens e hardware), mas pode ser diagnosticada e parcialmente enfrentada por práticas de ciência aberta, reuso de dados e software de código aberto. A crítica aqui é dirigida a tendência “nacionalista”³ que tratou a soberania como fechamento territorial ou

¹ Discuti em outra oportunidade as ambiguidades do conceito de “inteligência artificial”, destacando a ausência de uma definição amplamente aceita e sua dependência em relação aos paradigmas científicos e técnicos que orientam sua aplicação. Cf Ramon Rota (2026).

² Metadados são informações que descrevem, contextualizam e tornam recuperáveis outros dados: autoria, data, fonte, formato, licença, critérios de coleta, variáveis, versão e condições de uso. Em história digital, eles cumprem papel semelhante ao aparato crítico das fontes, pois permitem avaliar proveniência, cadeia de custódia, limites e possibilidade de reuso.

³ Por nacionalismo entende-se a naturalização da equivalência entre nação, Estado e sociedade como forma “normal” da modernidade, fazendo da moldura nacional o horizonte quase espontâneo da análise. A crítica não implica abandonar a nação como objeto histórico, mas problematizar sua conversão em unidade autoevidente da narrativa e da explicação histórica. Nesse sentido, a questão nacional aparece como problema transversal à historiografia: ela organiza modos de periodizar, territorializar e narrar o passado, além de definir sujeitos históricos, arquivos, escalas e categorias de análise. Cf Gopal Balakrishnan (2007).

autossuficiente: a história do conceito mostra que soberania é sempre disputada, relacional e atravessada por dependências externas.

A emergência do Debate

A popularização da Inteligência Artificial impôs a necessidade de reflexão sobre as consequências da IA generativa e dos grandes modelos de linguagem, tanto para sua fabricação quanto para a inferência (uso). Para o Aprendizado de Máquina Profundo aplicado a *Large Language Models* (LLMs), são necessários grandes volumes de dados e elevada capacidade de processamento, condições que incidem sobre a geopolítica internacional, à medida que bancos de dados e estruturas físicas de computação dependem de uma economia globalizada que distribui desigualmente seus dividendos entre as partes envolvidas.

No debate sobre o colonialismo de dados, argumenta-se que os dados pessoais, comportamentais e sociais se tornaram matéria-prima central do capitalismo contemporâneo, reproduzindo lógicas coloniais de extração, apropriação e assimetria de valor (CASSINO; SOUZA; SILVEIRA, 2022). Couldry e Mejias definem esse processo como uma nova fase de apropriação da vida social, na qual interações cotidianas são continuamente convertidas em dados e transformadas em recurso econômico e político para plataformas, corporações e Estados (COULDRY; MEJIAS, 2019). Para os autores, as infraestruturas digitais não superam as estruturas históricas de apropriação e desigualdade, mas as reconfiguram, produzindo uma nova ordem social baseada na captura, no processamento e na monetização da vida humana. Nesse sentido, o debate sobre colonialismo de dados aproxima-se da crítica do capitalismo e da economia política da informação, mas desloca o foco da exploração apenas do trabalho para formas mais amplas de extração e apropriação da vida social mediadas por dados.

Mas, se o capitalismo é um modo de produção não superado e o digital é *conditio sine qua non* de nossas vidas, é possível produzir uma história digital crítica de seus próprios limites? Como identificar, no digital, as assimetrias que a escrita da história procura combater?

A expressão “colonialismo de dados” já é utilizada por Thatcher, O’Sullivan e Mahmoudi (2016) em “Data Colonialism through Accumulation by Dispossession”. que define a captura assimétrica de dados como forma de acumulação por despossessão e commodificação da vida

cotidiana. O dado é produzido, agregado, interpretado e apropriado dentro de relações assimétricas de poder, assim, o *big data*, definido pelo volume, velocidade e variedade, perde seu efeito. Porque o que sustenta o conceito é o “valor” produzido a partir do relacionamento, não o ato de acumular e processar quantidades e diversidade com em menor tempo. Desse ponto de vista, a produção de *big data* tende a reproduzir lógicas de colonialismo de dados quando naturaliza a extração, a apropriação e a assimetria de valor que tornam possível sua própria existência.

Uma pista importante para compreender esse paradoxo aparece na obra de José van Dijck. Em vez de apresentar uma crítica do capitalismo de dados informada diretamente pela economia política e por debates pós-coloniais ou decoloniais, Van Dijck procura examinar os pressupostos epistemológicos e ontológicos que permitem converter modos de vida em dados e, posteriormente, em bancos de dados. Van Dijck (2014), ao mapear como a datificação se converteu em um paradigma ideológico, argumenta que a tolerância contemporânea à vigilância e ao acesso rotineiro a dados pessoais se explica pela normalização da datificação como paradigma social e científico. Nesse paradigma, transformar ações, relações e experiências em dados quantificados passa a parecer um modo legítimo de conhecer, administrar e governar o social. A pesquisadora sustenta que tal paradigma repousa numa ideologia, o “dataísmo”, marcada por pressupostos ontológicos e epistemológicos problemáticos, especialmente a crença de que os dados seriam objetivos, transparentes e capazes de permitir formas amplas de monitoramento, predição e intervenção social.

As constatações de Cassino *et al.*, Couldry, Mejias e van Dijck permitem identificar, de um lado, a dimensão econômico-política que a IA ganhou ao exponenciar disputas por dados, infraestrutura e capacidade de processamento e, de outro, a necessidade de produzir alternativas críticas e éticas ao dataísmo. Para a escrita da história sob uma ótica computacional, a expansão de grandes repositórios digitais e o uso de ferramentas orientadas a *big data* não significam automatização da técnica, mas integração crítica: incorporação de ferramentas computacionais sem abandono da teoria, da crítica documental e da contextualização historiográfica (OHARA, 2022; SILVEIRA, 2018, BONALDO, 2022).

Nessa chave, entre perspectivas digitais “apolíticas e integradas”, em referência a Umberto Eco, a crítica historiadora se reafirma como fundamento da escrita da história por meio do cruzamento de fontes internas e externas, da contextualização e da explicitação dos

procedimentos de prova. Para Tiago Gil (2024), ao confrontar *big data* e escrita da história, emerge um “neopositivismo digital”: a crença de que mais dados, maior escala e algoritmos seriam suficientes para substituir a interpretação histórica ou tornar dispensável a crítica das fontes. A crítica de Gil encontra na obra de Matthew L. Jockers (2013) uma de suas principais exemplificações: a conclusão de que a empiria, em grande escala, poderia superar a teoria.

Essa discussão não é inteiramente nova na historiografia. Em Documento/Monumento, Le Goff já indicava que a fonte documental é fabricada por operações sociais e técnicas, chegando a mencionar computadores e bancos de dados em escala muito distinta do *big data* contemporâneo (LE GOFF, 1990), mas em relação direta com as fundamentações computacionais que ainda mobilizamos. Portanto, se já não existe o risco de a prática historiadora converter linearmente documento em acontecimento, a escala das operações para fabricar história digital depende da exploração crítica das potencialidades dos recursos computacionais.

Autores clássicos da história digital como Cohen e Rosenzweig (2006) tiveram grande contribuição ao relacionar a preservação do digital com a prática historiadora, mas ainda são necessários avanços. Stephen Robertson (2022) argumenta que a história digital ainda precisa avançar para além da preservação de evidências produzidas na rede mundial de computadores. Assim, a pergunta não é apenas como preservar dados, mas como mobilizá-los como história quando o próprio meio computacional é distribuído de modo desigual no planeta. A resposta proposta aqui combina crítica historiográfica e estratégia comunitária: produzir história digital significa transformar operações de coleta, limpeza, classificação, agregação, visualização e preservação em procedimentos interpretáveis, documentados e discutíveis.

Na América Latina, essa questão ganha densidade específica porque a integração entre história e computação depende de recursos computacionais distribuídos de modo desigual: infraestrutura de armazenamento, capacidade de processamento, acesso a serviços de nuvem, padrões de metadados, repositórios públicos, formação técnica e comunidades capazes de manter ferramentas. Assim, a pergunta que orienta o artigo pode ser formulada de modo mais preciso: como produzir história digital crítica, em contexto latino-americano, diante do colonialismo de dados e da dependência de infraestruturas controladas por atores externos?

Uma história digital crítica, que reconhece possibilidades e limites dos algoritmos como parte da pesquisa, produz não apenas novas abordagens para a disciplina da História, mas

também alternativas soberanas quando praticada como ciência aberta e código aberto. Por alternativas soberanas compreendem-se práticas de pesquisa reutilizáveis no nível do código e dos bancos de dados, compatíveis com a infraestrutura disponível em universidades, laboratórios e repositórios públicos latino-americanos: rotinas auditáveis, replicáveis e comunitariamente negociadas que não dependam integralmente de treinamento local de modelos gigantescos nem de serviços opacos.

A discussão conceitual de soberania é necessário porque permite distinguir quais sentidos estão em jogo quando se fala em soberania digital. A decomposição de Krasner em dimensões legal-internacional, westfaliana, doméstica e de interdependência (KRASNER, 1999; 2001) ajuda a traduzir o problema para o digital: reconhecimento jurídico e regulação pública; proteção contra controle externo de infraestruturas; capacidade institucional de governar dados e arquivos; e controle relativo de fluxos transfronteiriços de informação, nuvem, custos e padrões técnicos. Desse modo, a discussão não é um adendo conceitual, mas o vocabulário que permite perguntar quem governa dados, plataformas, repositórios e evidências históricas.⁴

A leitura histórica do conceito permite superar clássicas colocações, como as de Hedley Bull, da Escola Inglesa de Relações Internacionais, para quem a política exterior funda uma sociedade internacional com instituições primárias para seu funcionamento, cujos princípios foram a diplomacia, a guerra, o direito internacional e o equilíbrio (BULL, 1977). Trata-se de uma visão revista pela historiografia das relações internacionais, de modo a sofisticar as “instituições primárias”, incluindo mudanças de autoridade e atores não estatais sem abandonar a gramática da soberania como instituição (BUZAN, 2004). Essa proposição permite identificar e reconhecer que a política internacional não é formada apenas pelos atores diretamente ligados

⁴ A crítica foucaultiana impede tratar a soberania apenas como atributo jurídico do Estado. Michel Foucault distingue o modelo do poder soberano — centrado no direito de fazer morrer e deixar viver — de formas modernas de poder disciplinar e biopolítico, voltadas à administração dos corpos, das populações e das condutas. Essa distinção não implica o desaparecimento da soberania, mas sua articulação com outros dispositivos de poder. Por isso, para este artigo, soberania digital deve ser lida também como disputa distribuída por infraestruturas, dados, normas, plataformas e regimes de governamentalidade (FOUCAULT, 2005). Além disso, Foucault é autor relevante nas discussões sobre vigilância, especialmente a partir de sua análise do panoptismo em *Vigiar e punir*, na qual a vigilância aparece como técnica de produção de sujeitos, normalização de comportamentos e organização assimétrica do visível. Essa contribuição permite compreender que, no ambiente digital, o problema da soberania não se limita ao controle jurídico sobre dados, mas envolve formas contínuas de observação, classificação, indução de condutas e produção de saberes sobre indivíduos e populações.

aos poderes oficiais que constituem a diplomacia dos Estados, mas também pela sociedade civil organizada.⁵

É interessante notar que a historiografia do conceito no Brasil identifica sua disputa e expansão de significado no Segundo Reinado. Conforme análise de Lauisa Rauter Pereira, a trajetória histórica do conceito de soberania (da tradição luso-brasileira de Antigo Regime até as disputas do Império) revela como a linguagem política se desagrega e se reconfigura quando a soberania deixa de ser um atributo estável do rei (legitimada por conquista, doação e ordem divina) e passa a ser atravessada pela ideia moderna de soberania do povo e da nação, o que torna o conceito um instrumento de disputa e abre dilemas práticos (representação, participação, partilha com o monarca, centralização vs. poderes provinciais) (PEREIRA, 2010).

Soberania é um conceito relativo e operativo, à medida que é invocado para normatizar disputas internacionais com diferentes pesos e produz consequências para os atores envolvidos. Tratando-se de um conceito relativo, como pode produzir novas alternativas? Sua instabilidade como dispositivo impõe restrições a escrita da história? Na próxima seção, procuremos colocar como a escrita da história pensou os relacionamentos entre Estados, mas principalmente como esta profissão contribui diretamente para os nacionalismos e sua crítica através da arqueologia de suas próprias práticas. Se essa arqueologia permite ver que se mantem os desafios da passagem para o físico ao digital, reconhece-se que há algo novo: os algoritmos. Mais que reconhecer sua existência e a implicação para nossa vidas, como utilizá-los de forma ética de forma a contribuir para a profissão? Quais alternativas existem para guardar a relação entre história e memória através do arquivo, agora digital?

História digital e algoritmos

Ao final da década de 1990, a história conectada apresenta-se como alternativa de escrita da história, inserida em um movimento historiográfico mais amplo de revisão das relações entre escrita da história e Estado-nação. Ao lado de abordagens como a história global,

⁵ Cabe destacar que a crítica ao nacionalismo na produção historiográfica foi realizada por diferentes movimentos historiográficos, principalmente pós-segunda guerra, não sendo, portanto, exclusividade da história das relações internacionais.

a história transnacional, a história cruzada, a história comparada e certas vertentes da história internacional, ela contribuiu para problematizar a tendência de tomar a nação ou o Estado como unidades naturais, autossuficientes e privilegiadas de explicação histórica. Essa crítica não implica abandonar a nação, o Estado ou o território como objetos de pesquisa, mas recusar sua conversão em molduras automáticas da narrativa histórica. Trata-se, portanto, de enfrentar o nacionalismo: a suposição de que os processos históricos podem ser adequadamente compreendidos a partir de fronteiras nacionais tomadas como evidentes.

Esse deslocamento de escala foi acompanhado por transformações mais amplas nas práticas que fabricam a história. A ampliação dos tipos documentais, a incorporação de arquivos dispersos, a atenção a fontes seriais, quantitativas, visuais, orais, materiais e digitais, bem como a valorização de sujeitos, circulações e experiências antes marginalizados, modificaram os modos de produzir evidência histórica. A historiografia passou a lidar, simultaneamente, com problemas de escassez e abundância: lacunas documentais, arquivos assimétricos, séries volumosas, fontes fragmentárias, diversidade linguística, variação de suportes e tensões entre quantidade e qualidade. Nesse sentido, a história conectada é relevante não apenas por propor outra escala de observação, mas por explicitar que as conexões históricas precisam ser construídas metodologicamente a partir de vestígios desiguais. (SUBRAHMANYAM, 2023).

Esses problemas reaparecem, de modo intensificado, quando se discute o *big data* na escrita da história. A história digital herda a preocupação com relações, circulações, escalas e assimetrias, mas só consegue mapear grandes volumes de conexões por meio de mediação computacional. Em bancos de dados, repositórios e plataformas, as conexões não aparecem espontaneamente: elas são produzidas por critérios de coleta, campos de metadados, filtros, chaves de relação, modelos de classificação, visualizações e algoritmos. Assim, o problema não é apenas utilizar computadores na pesquisa histórica, mas compreender como o computador participa da própria fabricação da evidência.

Essa mediação computacional introduz uma dimensão material incontornável. Softwares, bibliotecas, modelos e algoritmos operam nos limites impostos por processadores, memória, GPUs, armazenamento, energia, redes, custos de nuvem e regimes de acesso às plataformas. Portanto, as perguntas que podem ser formuladas, os bancos de dados que podem ser processados e os modelos que podem ser testados dependem das infraestruturas disponíveis em cada contexto institucional e geográfico. A passagem da “história tradicional”

produzida inclusive em ambientes computacionais para a história digital que explora suas potencialidades, não consiste, assim, em transpor literalmente um conceito historiográfico para o ambiente computacional, mas em levar adiante sua lição teórica-metodológica: escalas, vínculos e conexões são construídos, e essa construção precisa ser explorada e explicitada.

Desse ponto de vista, a questão deixa de ser apenas como usar algoritmos na escrita da história. Torna-se necessário perguntar quais algoritmos, dados, códigos, hardwares, plataformas e infraestruturas estão disponíveis; quem os controla; quais operações tornam possíveis; e que formas de dependência produzem. É nesse ponto que a história digital se articula ao problema do colonialismo de dados e da soberania digital: ela ajuda a pensar as clássicas relações e assimetrias de poder, enquanto a crítica computacional da escrita da história permite compreender como essas relações são tecnicamente produzidas, limitadas ou obscurecidas no ambiente digital. Por outro lado, considerar a historicidade do digital e da escrita da escrita conduz a reavaliar argumentações nacionalistas e autossuficientes, sobretudo também porque que a noção de soberania é paradoxal e relativa. Se de um lado é bem verdade que os produtos tecnológicos são produzidos e explorados de maneira desigual, de outro, sua produção é produto do coletivo, com dados compartilhados. Não estamos distantes da sentença de Walter Benjamin, para quem “Nunca houve um monumento da cultura que não fosse também um monumento da barbárie”.

Como alternativa aos modelos proprietários e poucos transparentes, o software livre o código aberto são as melhores alternativas para contextos latinos-americanos, sobretudo porque permitem desenvolvimento intelectual e adaptação da computacional a realidade do lugar, mas sem cair um em realismo historiográfico tampouco nacionalista.

A conexão algorítmica é, nesse sentido, uma abordagem relacional para o suporte digital: ela designa rotinas que filtram, agregam, ordenam, classificam, cruzam e visualizam vestígios em bancos de dados. Algoritmo é entendido aqui como uma sequência formalizada de instruções que transforma entradas em saídas por meio de regras de classificação, ordenação, filtragem ou cálculo; por isso, pode ser mobilizado como método quando organiza procedimentos reprodutíveis de pesquisa e como problema teórico quando suas escolhas produzem visibilidade, exclusão e evidência (O'NEIL, 2016). Portanto, se os algoritmos são ambivalentes, cabe a história digital desenvolver suas próprias recursividades algorítmicas como retorno crítico sobre essas próprias operações: os resultados gerados por rotinas

computacionais devem ser reavaliados historiograficamente e explicitadas as condições técnicas de produção da evidência.

O algoritmo é o instrumento que operacionaliza essa analogia quando produz mapas de relações, séries temporais, agrupamentos, filtros e visualizações, desde que sua elaboração seja documentada e criticada pela escrita da história. A diferença entre história digital e usos genéricos da computação nas humanidades reside no sistema de referência: a evidência precisa permanecer lastreada em fontes, metadados, código, decisões de coleta e cadeia de custódia, sem que a “automação” substitua a interpretação.

Para a história digital, a evidência não se sustenta apenas pela imagem ou pelo recorte extraído de uma plataforma: ela depende de informações complementares sobre autoria, coleta, temporalidade, circulação, documentação técnica e contexto historiográfico, que podem ser articuladas pela bibliografia especializada ou por *trending topics*⁶. Porém, um *print* de uma rede social ou a projeção gráfica gerada a partir de tecnologias de plataformas de *Big Techs*⁷ oferece poucos recursos para essa contextualização do ponto de vista digital, porque tais abordagens não capturam o funcionamento dos algoritmos que tornaram possíveis tanto a publicação quanto a representação gráfica. Nesses casos, contextualizar a evidência exige combinar dados externos às plataformas com dados produzidos pelo próprio processo de coleta e tratamento, sem perder a cadeia de custódia que tornou a evidência destacada possível.

Dedicar maior atenção às bases de dados não significa um retorno à empiria ou ao realismo histórico, mas a necessidade de aprimorar a história digital ao tornar visíveis as operações que produzem a evidência: recorte, coleta, limpeza, classificação, infraestrutura, plataforma, metadados e visualização. Esse é o elo que conduz aos cinco recortes seguintes, pois cada um deles mostra uma camada diferente do problema do colonialismo de dados e da soberania digital.

⁶ Trending topics são listas, rankings ou indicadores de temas em circulação intensa em plataformas digitais, produzidos por métricas e algoritmos próprios de cada serviço. Para a escrita da história, podem funcionar como indícios de circulação e temporalidade pública, mas exigem crítica porque resultam de critérios opacos de coleta, ordenação e recomendação.

⁷ *Big Techs* designa grandes corporações de tecnologia que concentram infraestrutura, plataformas, dados, serviços de nuvem, publicidade, sistemas operacionais ou ecossistemas de desenvolvimento.

Recortes a partir de bancos de dados

A combinação de linguagens computacionais para elaboração de rotinas algorítmicas será demonstrada a partir de cinco recortes baseados em bancos de dados, tratados simultaneamente como suporte de evidência e como objetos de crítica. Eles não funcionam apenas como exemplos descritivos, mas respondem às hipóteses anunciadas no início do texto. O primeiro mostra a dependência contratual de serviços privados; o segundo, a concentração geopolítica da capacidade de IA; o terceiro, a materialidade extrativa da infraestrutura digital; o quarto, as assimetrias de visibilidade da produção latino-americana; e o quinto, uma demonstração de reuso de dados com crítica de fonte, metadados e preservação.

Tabela 1: Conjuntos de dados/fontes digitais e função analítica

Conjunto de dados/fonte	Origem	Período	O que mede	Como foi processado
Contratos públicos com Big Techs	Silva <i>et al.</i> (2025) (base em dados públicos)	s/d (conforme recorte do estudo)	Gastos/contratos com serviços digitais (nuvem, licenças etc.)	Reuso de agregações e categorias; crítica do recorte por palavras-chave
AI Compute Sovereignty	Hawkins; Lehdonvirta; Wu (2025)	snapshot 2025	Distribuição de regiões de treinamento e inferência (cloud/IA)	Replot/visualização com Python (Pandas/Matplotlib) + interpretação
Digital Economy Report 2024	United Nations (2024) UNCTAD	2019–2021	Exportações de minerais/minérios /metais (participação no total)	Leitura e contextualização do indicador como cadeia material da IA
Mapa institucional (SciELO) – universidade com produção em	Elaboração do autor (metadados SciELO + afiliações institucionais)	desde 2000 (coleta até jan. 2026)	Distribuição geográfica de universidades latino-americanas com produção identificada em história digital e	Extração de metadados, normalização institucional, geocodificação e clusterização/visualização

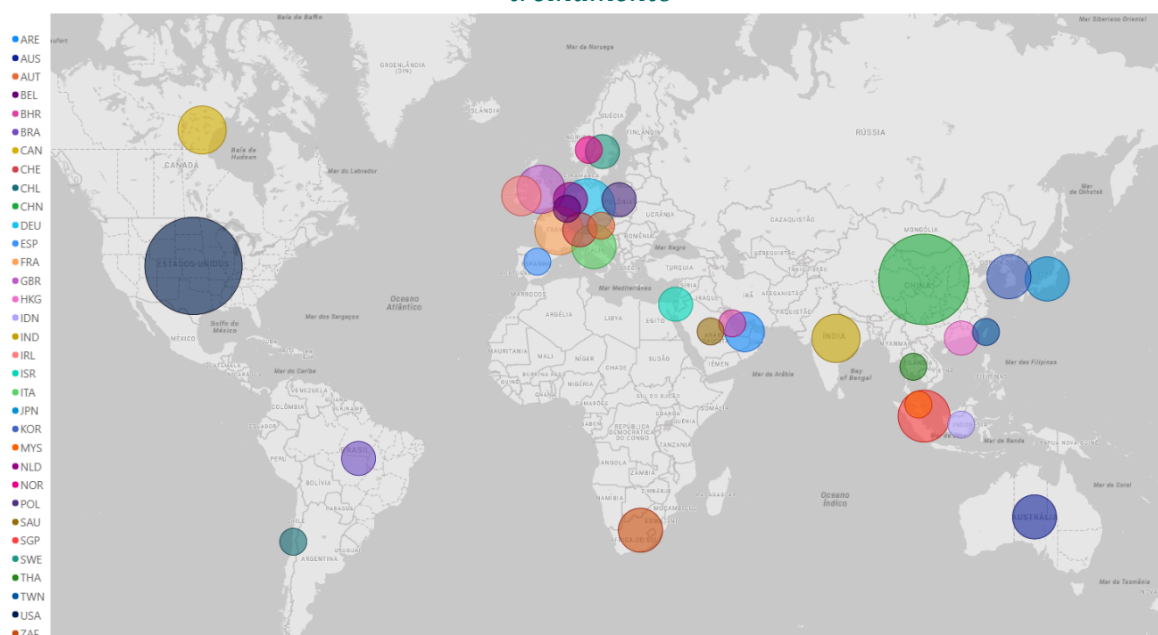
Conjunto de dados/fonte	Origem	Período	O que mede	Como foi processado
história digital/HD			humanidades digitais	
Uma década conectada (X)	Rota <i>et al.</i> (2023) (ReDU/Unicamp)	2007–2023	Volume/ritmos de menções	Normalização temporal, agregação diária e visualização; recortes por termos

Fonte: elaborado pelo autor.

No primeiro recorte, tomo como ponto de partida o estudo de Ergon Cugler de Moraes Silva *et al.* (2025) que, a partir da análise de bases públicas sobre contratos estabelecidos pelo governo federal, revela um orçamento de 23 bilhões para aquisição de serviços computacionais, principalmente de empresas como Amazon, Oracle, Microsoft, Google, Meta, Apple e Red Hat, distribuídos entre serviços de nuvem, licenças e segurança. O banco de dados em questão foi construído a partir de palavras-chave que recortaram as bases originais como método de seleção do conteúdo. Além do mapeamento, as autoras e os autores representaram o que poderiam significar os 23 bilhões se fossem investidos em infraestrutura e formação profissional, parte de um ciclo a ser consolidado como política de Estado para a construção de soberania nacional. A metáfora explorada no estudo é exemplar: o Estado brasileiro decidiu por um contrato *ad aeternum* de aluguel. Este recorte demonstra que a dependência infraestrutural é produzida e estabilizada por contratos e regimes privados de acesso. Analiticamente, esse recorte mostra que o colonialismo de dado, para a coleta de informações, depende de contratos por aluguel contínuo de infraestrutura crítica.

No segundo recorte, utilizo o mapeamento realizado por pesquisadores da Universidade de Oxford para compreender o treinamento de modelos de inteligência artificial de acordo com os Estados nacionais, representado na figura abaixo. Esse recorte desloca o problema da soberania digital do plano abstrato para a geografia concreta da computação: onde estão as regiões de treinamento e inferência e quem controla seu acesso.

Figura 1: Representação da distribuição global das infraestruturas de aprendizado de máquina voltadas para grandes modelos de linguagem. Os círculos são proporcionais à capacidade de treinamento



Fonte: Adaptado a partir dos dados de Hawkins, Zoe and Lehdonvirta, Vili and Wu, Boxi, AI Compute Sovereignty: Infrastructure Control Across Territories, Cloud, 2025.

A figura acima foi produzida a partir das tabelas apresentadas por Hawkins et al (2025) para o mapeamento de uso e inferência de Inteligência Artificial. Utilizou-se as categorias definidas no estudo, como país e quantidade de uso e inferência para gerar uma representação da distribuição dos valores com linguagem *Python* e as bibliotecas *Pandas* e *Matplotlib*.⁸ De um estudo quantitativo abstraiu-se categorias qualitativas para enfatizar a dependência de hardware para produção de Inteligência Artificial. Na Figura 1, a distribuição das regiões de computação em nuvem voltadas à IA deixa nítido que o gargalo central não são os saberes para treinar modelos — já dominado por universidades, grandes laboratórios, sociedade organizada e empresas de programas computacionais em dezenas de países —, mas sim a infraestrutura física de data centers com processamento de última geração. A tabela evidencia três grandes blocos:

⁸ Python é uma linguagem de programação de alto nível, amplamente usada em ciência de dados por permitir rotinas reprodutíveis de coleta, limpeza, cálculo e visualização. Bibliotecas são conjuntos de funções prontas que ampliam a linguagem: Pandas organiza bases em tabelas, filtra, agrega e calcula grandes volumes de dados; Matplotlib produz gráficos e visualizações. Esses recursos permitem a prática historiadora explicitar procedimentos, testar recortes e apresentar evidências em escala.

Tabela 2: Principais países em treinamento e inferência

Bloco	Perfil	Países-exemplo	Critério quantitativo
1. Hegemônicos	Clusters locais robustos para treinar e servir modelos	China (22/4/22), EUA (26/22/13), Alemanha (7/4/4), França (5/3/3), Canadá (5/3/2), Índia (5/3/3), Reino Unido (5/3/3), Japão (4/4/1), Austrália (4/3/1), Coreia (4/3/2)	≥ 4 regiões e ≥ 3 dedicadas a <i>training</i>
2. Capacidade mista	Infra moderada; treinam apenas modelos de médio porte	Bélgica, Brasil, Irlanda, Israel, Países Baixos, Noruega, Polónia, Singapura (6/2/5), África do Sul, Espanha, Suécia, Suíça, Emirados Árabes	1 — 4 regiões totais e ≥ 1 para <i>training</i>
3. Inferência-first	Só executam ou ajustam modelos prontos; <i>training</i> inexistente	Áustria, Bahrain, Chile, Hong Kong, Indonésia, Malásia, Arábia Saudita, Taiwan, Tailândia	1 — 2 regiões totais, 0 para <i>training</i> , ≥ 1 para <i>inference</i>

Fonte: Hawkins, Zoe and Lehdonvirta, Vili and Wu, Boxi, AI Compute Sovereignty: Infrastructure Control Across Territories, Cloud, 2025.

Com 26 regiões totais e 22 de treinamento, os EUA superam todos os demais países em capacidade — ficando atrás apenas da China no número de regiões dedicadas à inferência. A própria China mantém 22 regiões de inferência e 4 de treinamento, superando toda a “capacidade mista” latino-americana somada. Mesmo na União Europeia, países como Bélgica, Áustria e Países Baixos dispõem de mão de obra qualificada, mas dependem de data centers distribuídos pela Irlanda, Alemanha ou França — ou de créditos em clouds externas. Sem infraestrutura local, a governança de dados sensíveis (saúde, defesa, energia) fica condicionada aos Acordos de Nível de Serviço e às políticas dos provedores estrangeiros, sobretudo para nações que não contam com a proteção de blocos multilaterais — dentro ou fora da UE, ou mesmo dos recém-estruturados Brics (sigla para o acordo bilateral entre países em desenvolvimento iniciado por África do Sul, Brasil, Rússia, Índia e China).

No curto prazo, essa dependência acelera a adoção de IA via Interfaces de Programação de Aplicações (API);⁹ no longo prazo, aprofunda a assimetria de poder computacional. Some-se a isso a dificuldade de obter métricas públicas confiáveis sobre clusters de treinamento. O estudo em destaque conseguiu discriminar 27 países com dados de treinamento, em apenas 10 encontraram informações consistentes sobre inferência — reforçando a opacidade do cenário global de infraestrutura de IA.

Por outro lado, em estudos que buscam mapear padrões globais, é recorrente a sub-representação de países em desenvolvimento, não necessariamente por ausência de dados, mas por escolhas de indexação, critérios de coleta e barreiras linguísticas. Países latino-americanos, em especial o Brasil, têm tradição forte em políticas de ciência aberta, mas nem sempre publicam nos mesmos circuitos e repositórios usados como base para mapeamentos internacionais. Esse ponto importa para a soberania digital: diagnósticos de “ausência” podem invisibilizar iniciativas locais de desenvolvimento de modelos específicos, reuso de modelos de código aberto e arranjos acadêmicos de infraestrutura. Assim, este segundo recorte sustenta que a capacidade computacional é concentrada e pouco transparente, e dialoga com o próximo recorte ao mostrar que assimetrias de visibilidade também integram a disputa por soberania digital.

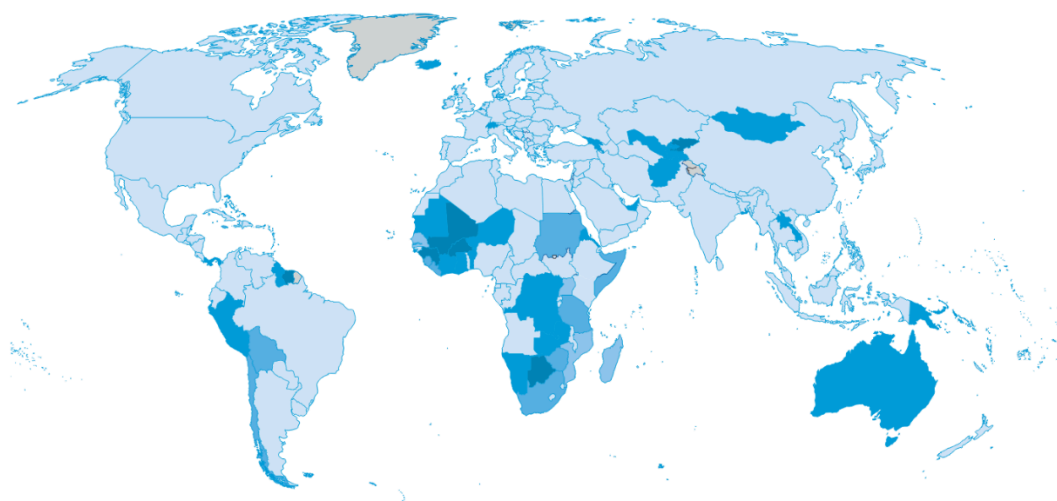
Quando observamos a lógica de contratos esboçados na primeiro reuso de dados e com o segundo, percebe-se que os agentes no exercício do poder tem poder de decisão capaz de fortalecer ou fragilizar alternativas soberanas. Nesse sentido, as políticas de dados abertos, código livre, participação comunitária tendem a ser alternativas viáveis para produção de saberes algoritmos, contornando até mesmo a dificuldade em produzir hardware livre, ou de distribuir igualmente os dividendos que são produzidos na sua comercialização.

No terceiro recorte, conecto os diagnósticos infraestruturais a indicadores internacionais sobre economia digital e extrativismo. Os dois estudos supracitados permitem concluir que a América Latina é extremamente dependente da infraestrutura internacional para compor sua infraestrutura nacional do ponto de vista computacional, em parte pelas características da globalização econômica que criou diferentes formas de agência para exploração de recursos naturais e criação de mercados consumidores. Conforme mapeamento

⁹ API, do inglês *Application Programming Interface*, é uma interface que permite a comunicação entre sistemas. No caso de serviços de IA, o usuário envia solicitações a modelos hospedados por terceiros, sem controlar integralmente infraestrutura, parâmetros, custos, logs, dados de entrada ou condições de permanência do serviço.

apresentado no Digital Economy Report 2024, embora exista um mercado global, a interdependência das economias cria ônus e bônus desiguais para cada Estado, tanto nas questões ambientais como na divisão dos lucros a partir da exploração de recursos naturais e valorização do trabalho (UNITED NATIONS, 2024). O terceiro recorte, portanto, dá suporte à crítica ao colonialismo de dados, à cadeia material que sustenta IA, nuvem e processamento: mineração, energia, logística, trabalho e mercado.

Figura 2: Participação percentual de minerais, minérios e metais no total de exportações de mercadorias, 2019–2021; tons mais escuros de azul indicam participação mais elevada



Fonte: UNITED NATIONS (2024). *Digital Economy Report 2024: Shaping an Environmentally Sustainable and Inclusive Digital Future*. Nova York: United Nations.

A Figura 2 representa a participação percentual das exportações de minerais, minérios e metais no total das exportações de mercadorias por país, no período de 2019 a 2021, isto é, o peso das commodities minerais na pauta exportadora. Na escala cromática, quanto mais escuro o azul, maior a participação percentual dos minerais, minérios e metais no conjunto das exportações do país. A dependência muito alta (60–100%) é visível em vários países da África subsaariana, como República Democrática do Congo, Mauritânia, Guiné e Zâmbia. Na América Latina, Chile e Peru apresentam participação elevada (entre 40–80%), associada sobretudo ao cobre. Para a economia política da IA, o ponto relevante é que a materialidade da infraestrutura digital se distribui desigualmente e reatualiza assimetrias históricas do capitalismo global. Este recorte subsidia a terceira hipótese do artigo, ao mostrar que a dependência computacional se

ancora em cadeias globais de valor que conectam mineração, energia, logística e infraestrutura digital.

Somando as questões geopolíticas e das relações internacionais que tornam a popularização da IA possível, ela traz implicações para diferentes contextos de cada país. No Brasil, o aprendizado de máquina aplicado a grandes modelos de linguagem impõe tensões às leis de direitos autorais, articuladas à Lei do Arquivo, Lei de Acesso à Informação, Marco Civil da Internet e Lei Geral de Proteção de Dados.¹⁰

O PL n. 2.338/2023, aprovado no Senado e encaminhado à Câmara dos Deputados, dialoga com pontos já consolidados em marcos regulatórios internacionais, como o AI Act da União Europeia, aprovado em 2024, que organiza obrigações e restrições a partir de uma abordagem por níveis de risco, com exigências de transparência, governança e documentação (COMISSÃO EUROPEIA, 2024). No Brasil, os esboços dos artigos 28 e 29 do PL n. 2.338/2023 versam sobre proteção de direitos autorais, ampliados para mídia em geral (texto, imagem, gravações de voz etc.), definindo consentimento e remuneração para uso de obra em treinamento. Para casos experimentais, prevê “balões de ensaio” regulamentados para testes em laboratório, com eventual retreinamento com dados autorizados quando houver finalidade comercial (SENADO FEDERAL, 2023).¹¹

A lei é importante, mas precisa definir também o papel das Big Techs que já utilizaram obras com direitos autorais e atualmente faturam com conteúdo gerado. Ainda há lacunas, como no uso de reconhecimento facial e na imprescindível presença de fiscalização durante todo o processo de treinamento por uma questão técnica: o aprendizado de máquina profundo para produção de LLMs dilui qualquer rastreamento de “citações” conforme conhecemos. Isso pode criar um imbróglio jurídico porque a identificação das obras usadas no treinamento depende das bases originais e de documentação de coleta, enquanto as saídas do modelo permitem apenas inferências probabilísticas sobre semelhança, influência ou reprodução.

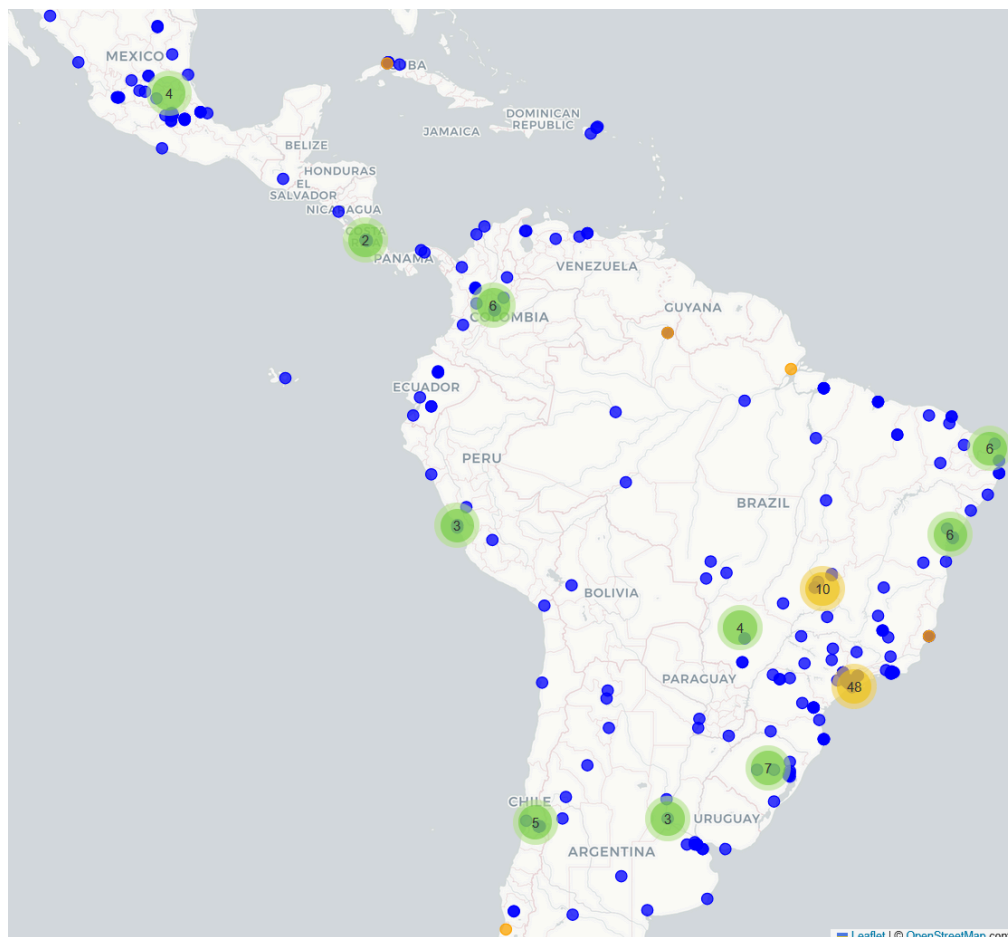
¹⁰ A Lei de Arquivos (Lei n. 8.159/1991) organiza a política nacional de arquivos públicos e privados; a Lei de Acesso à Informação (Lei n. 12.527/2011) regula transparência e acesso a informações públicas; o Marco Civil da Internet (Lei n. 12.965/2014) estabelece direitos, deveres e garantias para o uso da internet; e a Lei Geral de Proteção de Dados (Lei n. 13.709/2018) regula o tratamento de dados pessoais.

¹¹ O PL n. 2.338/2023 é a proposta brasileira de marco regulatório para sistemas de inteligência artificial. O Senado aprovou substitutivo em 10 dez. 2024 e encaminhou autógrafo à Câmara dos Deputados em 17 mar. 2025, mantendo o debate legislativo sobre riscos, direitos, transparência, governança e responsabilidade.

Diante da necessidade de Estados nacionais produzirem instrumentos para sua proteção e, ao mesmo tempo, operarem em um sistema internacional assimétrico, a soberania reaparece como necessidade paradoxal. Os quatro primeiros recortes mostram essa tensão em níveis distintos: contratos públicos, infraestrutura de IA, cadeias extrativas e visibilidade acadêmica. O quinto recorte recebe seção própria porque funciona como demonstração metodológica: nele o artigo mostra, em escala menor e auditável, como dados podem ser reutilizados historiograficamente sem abandonar crítica de fontes, documentação e preservação.

No quarto recorte, apresentamos um banco de dados original construído a partir do cruzamento entre metadados de periódicos científicos e afiliações institucionais, com o objetivo de identificar e georreferenciar universidades latino-americanas que publicaram, desde 2000, sobre história digital e humanidades digitais (base *SciELO Crossref* e *OpenAlex*). O propósito do levantamento não é tomar volume como sinônimo de maturidade do campo, mas tornar visíveis práticas de institucionalização e circulação do conhecimento (grupos, laboratórios, redes e periódicos) que operam tanto como crítica da datificação quanto como apropriação reflexiva de abordagens computacionais. Este recorte evidencia como ocorrem assimetrias de visibilidade no digital. Analiticamente, este quarto recorte não mede “maturidade” do campo, mas evidência como regimes de indexação, idioma, metadados e afiliação produzem visibilidades e invisibilidades na história digital latino-americana.

Figura 3: Distribuição georreferenciada de universidades latino-americanas com produção identificada em história digital e humanidades digitais (bases SciELO, Crossref e OpenAlex), desde 2000



Fonte: elaboração do autor, a partir de metadados SciELO, Crossref e OpenAlex e afiliações institucionais; cada ponto indica uma universidade georreferenciada. Os números nos agrupamentos (*clusters*) indicam a quantidade de instituições/publicações agregadas naquela área visível da escala do mapa; as cores diferenciam a densidade dos agrupamentos.

A Figura 3 evidencia uma distribuição desigual de instituições, com maior densidade em torno de grandes sistemas universitários e áreas metropolitanas, destacando-se o Sudeste do Brasil e o México, seguidos por nós relevantes no eixo andino e no Sul, além de uma dispersão que sugere circulação e cooperação para além das capitais.

Esse padrão reflete assimetrias materiais (financiamento, conectividade, infraestrutura e regimes de indexação) dentro da América Latina, mas indica também produção intelectual relevante não considerada. Nesse sentido, se “colonialismo de dados” nomeia circuitos de

extração, assimetria de valor e concentração de infraestrutura, o próprio “dados” que sustenta a crítica colonial pode estar incompleto ou sub representado. Assim, a crítica do colonialismo de dados precisa ser elaborada com cuidado para não apagar gradações e alternativas. O mapa aponta, antes, para “soberanias parciais” e incrementais, observáveis em práticas de publicação, curadoria e preservação, ainda que sob um sistema internacional assimétrico.

Dados de reuso: do quantitativo ao qualitativo

Nesta última seção, discuto o reuso de dados a partir de um dataset de publicações da plataforma X arquivado em repositório acadêmico, articulando ciência aberta, preservação e crítica de fontes digitais. Ele recebe uma seção própria porque deixa de ser apenas diagnóstico e passa a demonstrar o segundo procedimento metodológico defendido pelo artigo: reutilizar dados, explicitar metadados, documentar recortes e interpretar historicamente rastros de plataforma. Assim, a Ciência Aberta é decisiva para a produção científica na América Latina porque cria as condições institucionais e técnicas para que dados, métodos e resultados circulem com transparência, auditabilidade e maior impacto social. Para explicitar por que esse reuso de dados se vincula a políticas soberanas, discutirei os conceitos de *commons* (os “comuns”), software livre/código aberto e ciência aberta.

Cada um desses campos possuem histórias próprias de desenvolvimento, mas se relacionam pela mesma arquitetura de problemas: o que pode ser compartilhado e sob quais condições; quais infraestruturas jurídico-técnicas tornam o compartilhamento operável (licenças, padrões, formatos, identificadores persistentes, repositórios); e quais formas de governança sustentam o uso legítimo e durável desses recursos (curadoria, responsabilização, critérios de justiça, sustentabilidade). Nessa chave, a contribuição para a soberania digital não está apenas na autoria individual, mas em instituir capacidades públicas de decidir sobre regras, infraestruturas e evidências digitais, evitando dependência estrutural de plataformas opacas e reduzindo assimetrias de controle sobre dados.

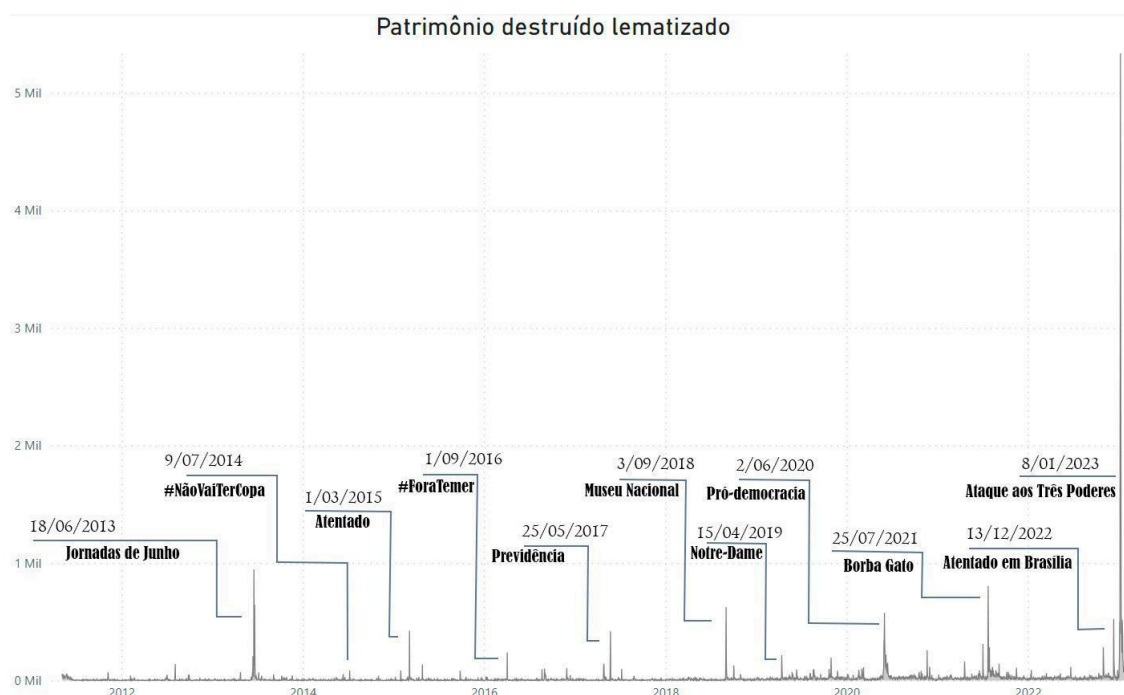
A literatura clássica sobre os *comuns* ajuda a evitar um equívoco recorrente: *commons* não é sinônimo de “acesso irrestrito”. Hardin formulou o dilema do uso coletivo quando não há mecanismos de coordenação e responsabilização (HARDIN, 1968); em contraponto, Ostrom

mostrou que arranjos policêntricos e regras bem desenhadas podem viabilizar autogovernança e sustentabilidade no uso de recursos compartilhados (OSTROM, 2009). Quando deslocamos esse debate para o conhecimento e o digital, o ponto decisivo é que a abertura só produz valor público se vier acompanhada de instituições e direitos de uso que organizem a produção, a preservação e a crítica. É nesse sentido que “o comum” pode ser lido como práxis instituinte: criar dispositivos, regras e garantias de uso que não se deixam reduzir à dicotomia Estado/mercado e que dão lastro normativo à ação coletiva (LAVAL; DARDOT, 2017).

No plano técnico, o software livre e o código aberto fornecem mecanismos concretos para transformar infraestrutura digital em um “comum” tratável: licenças e acesso ao código tornam possível auditar, reproduzir, adaptar e manter ferramentas, códigos, repositórios, versionamentos ao longo do tempo. Benkler descreve esse sistema como um modo de produção baseado em colaboração distribuída e governança por regras compartilhadas, com implicações diretas para como ciência e cultura se organizam em ambientes digitais (BENKLER, 2002). A ciência aberta, por sua vez, amplia esse horizonte para a universidade e para a pesquisa (dados, software, métodos, avaliação), convertendo princípios em políticas e métricas de monitoramento que pressionam por transparência, reuso e responsabilização institucional (BOBROV *et al.*, 2025). Assim, quando universidades sustentam repositórios, padrões de metadados, curadoria e licenças, elas não apenas “publicam dados”: elas produzem soberania digital ao garantir que evidências permaneçam verificáveis, preservados e governados publicamente, inclusive quando as plataformas de origem (como a X) restringem acesso, removem conteúdo ou alteram regras de disponibilidade (ROTA; NICODEMO, 2024).

O gráfico abaixo explora dados coletados a partir de publicações que mencionaram a palavra “patrimônio”, entre os anos de 2007 e 2023, período que permite relacionar a criação da rede social Twitter (atual X), com o atentado à sede dos Três Poderes em Brasília. Porém, os dados estratosféricos impedem representar algo além de um número astronômico. Entre as publicações que mencionaram “patrimônio”, recortou-se uma categoria específica: destruídos, vandalizados e ressignificados. Observe que na Figura 4 há dois grandes picos sobre os patrimônios: em 2013, nas Jornadas de Junho e 2023, no Ataque às Sedes dos Três Poderes. Entre elas diversas discussões sobre acontecimentos que modificaram o entendimento do debate público sobre o que é patrimônio.

Figura 4: Representação das publicações em português divididas por dia que mencionaram ao menos as palavras *patrimônios destruídos*, de forma lematizada¹²



Fonte: ROTA, Alesson Ramon *et al.* (2023). Dataset: uma década conectada: patrimônios incômodos no paradigma 2013–2023. Versão 2.0. Campinas: Universidade Estadual de Campinas, Repositório de Dados de Pesquisa da Unicamp, 2023. Disponível em: <https://doi.org/10.25824/redu/IAQJRJ>. Acesso em: 30 jan. 2026.

A coleta em questão foi um estudo experimental para compreender a noção ampliada de patrimônio, isto é, como ela foi mobilizada pelo público, em vez de partir das noções consolidadas pela bibliografia especializada. Para o estudo em questão, foi possível enxergar o deslocamento do uso do patrimônio como espaço público em direção ao patrimônio histórico na rede social. (ROTA e Nicodemo, 2024) Como pano de fundo dessa cadeia de sentido identificada, estão acontecimentos históricos que forçaram o debate sobre o patrimônio, mas também iniciativas de instituições especializadas na preservação e educação patrimonial.

As conexões entre 2013 e 2023, porém, não se limitam às questões patrimoniais, que são apenas um recorte algorítmico para o banco de dados, mas com interpretações que buscam compreender os acontecimentos históricos recentes. As Jornadas de Junho de 2013 foram

¹² Lematização é uma técnica de processamento de linguagem natural que reduz formas flexionadas de uma palavra ao seu lema ou forma de dicionário. Assim, variações como singular/plural, feminino/masculino ou conjugações verbais podem ser reunidas para ampliar a recuperação e comparar ocorrências sem tratar cada flexão como termo isolado.

movimentos sociais legítimos, em um primeiro momento, que reivindicavam melhores condições de vida. Na época, popularizou-se o Catraca Livre e foi criada a Mídia Ninja para acompanhar os protestos. Foi justamente esse segundo veículo que gerou a curva no gráfico acima, no dia 18 de junho de 2013, com a divulgação de mídias alternativas para o trabalho de imprensa. Contudo, os movimentos passaram a ser instrumentalizados com a repercussão dos *black blocs*. Os protestos por passe livre — desencadeados pelo aumento nas passagens de ônibus quando o Brasil crescia mais que o Japão e oscilava entre a sexta e a sétima economia do mundo — ganharam outros contornos a partir da organização do Movimento Brasil Livre (MBL) e do Vem Pra Rua (VPR) (MOURÃO; LEPRI, 2020; COTA *et al.*, 2019; JARDELINO *et al.*, 2020).

A Figura 4 mostra que o reuso de dados de plataforma depende das conexões estabelecidas pela rede mundial de computadores, mas o ponto central não é mapear a web em si. O que importa para a argumentação é que as plataformas organizam a circulação por operações algorítmicas que selecionam, ordenam e tornam visíveis certos registros. Quando esses resíduos são preservados em bancos de dados e tratados criticamente, a história digital pode explicitar sistemas de referência, preservação e autoria, sem tomar a automação como evidência autônoma.

Para a produção da Figura 4, foram programadas rotinas algorítmicas em *Python* para analisar a base de dados disponibilizada no Repositório de Dados de Pesquisa da Unicamp (ReDU), composta por publicações coletadas na plataforma X. Após a ingestão do conjunto (com a biblioteca de programação *Pandas*), as postagens foram normalizadas temporalmente (conversão de datas e padronização do fuso), agregadas por dia (*groupby* da data¹³) e convertidas em uma série temporal com o total diário de publicações, que então foi visualizada como linha contínua em uma linha do tempo (via *Pandas* e *Matplotlib*). O gráfico apresenta um quadro geral do volume e da variação das menções (no caso, associadas a “patrimônio destruído” em forma lematizada), mas a mesma rotina permite aplicar filtros — por termos, categorias, janelas de tempo ou subconjuntos de eventos — para isolar curvas específicas e comparar padrões de recorrência, picos e períodos de estabilidade. Ao tratar o banco de dados como suporte de evidência, essa abordagem torna observável como certas informações

¹³ *groupby* é uma funcionalidade da biblioteca *Pandas* que agrupa linhas de uma tabela segundo uma ou mais variáveis, como datas, autores, categorias ou países, permitindo calcular totais, médias, frequências e outros agregados. No exemplo, as publicações foram agrupadas por dia para gerar uma série temporal.

repercutem publicamente, em que intensidade, e com que ritmos, articulando circulação discursiva e temporalidades de engajamento a partir de rastros computacionais.

Nesse procedimento, Python, Pandas e Matplotlib não são soberanas por si mesmos, mas exemplos de uma infraestrutura sociotécnica aberta que torna a soberania pensável a partir de práticas comunitárias. Seu valor não decorre apenas de serem ferramentas gratuitas, mas do modo como são mantidas: documentação pública, revisão coletiva, negociação de padrões, correção distribuída de erros e possibilidade de auditoria, adaptação e reuso. A fronteira entre usuário e produtor torna-se mais porosa, pois pesquisadores, desenvolvedores, instituições e comunidades podem discutir rotinas, corrigir bibliotecas, publicar códigos e tornar visíveis os critérios de tratamento dos dados. Essa lógica comunitária oferece uma defesa parcial contra formas de dominação associadas ao colonialismo de dados, porque reduz a dependência de plataformas proprietárias opacas e amplia a capacidade local de produzir, verificar e preservar evidências digitais.

Esse exercício de mensurar a repercussão de dados em linhas do tempo, aplicando diferentes recortes e camadas de filtragem, aproxima-se da história dos conceitos formulada por Koselleck (1992), na medida em que busca tornar visíveis tanto as diacronias quanto as sincronias da linguagem em circulação. Se, na tradição koselleckiana, a análise conceitual procura captar deslocamentos de sentido ao longo do tempo e, simultaneamente, as disputas semânticas que coexistem em um mesmo momento histórico, a exploração de séries temporais derivadas de bancos de dados permite mapear essas variações em outra escala e suporte: a dos rastros digitais. Assim, os picos, inflexões e recorrências não são tomados como equivalentes ao “sentido” em si, mas como indícios da intensidade, da oportunidade e da reconfiguração pública de vocabulários e enquadramentos, que podem ser reconstituídos por meio da leitura contextual dos eventos e de suas controvérsias.¹⁴ Desse modo, a análise computacional não substitui a interpretação histórica; ao contrário, ela oferece um dispositivo de observação para orientar a crítica das fontes, formular hipóteses sobre mudanças semânticas e explicitar como, em ambientes plataformizados, a circulação linguística se torna também um problema de infraestrutura e mediação algorítmica

¹⁴ É importante salientar que uma história digital não conduz obrigatoriamente para a história conceitual. É plausível que todas as teorias da história possam ser relacionadas com o digital dado que é um fenômeno social, por isso, é aconselhável compreendê-lo a partir de suas características. Cf Peter Bola (2023) que define a escrita da história no digital como uma forma de se relacionar com palavras dinâmicas, para, a partir delas, acessar o sentido histórico.

Os rastros de plataforma não equivalem ao "debate público" em sua totalidade. A composição do *dataset* é condicionada por critérios de coleta, disponibilidade e eventuais remoções, além de assimetrias de participação, presença de automação (*bots*), mudanças de regras da plataforma e ruídos de linguagem (ambiguidade de termos e efeitos da lematização). Por isso, picos e variações aqui apresentados devem ser lidos como indícios de circulação e intensidade, sempre dependentes de crítica historiográfica. A análise prioriza agregações e não tem como objetivo identificar indivíduos, reduzindo riscos de exposição. O reuso de dados demanda transparência sobre origem, licenças, documentação e limites do compartilhamento, bem como atenção a marcos como proteção de dados e direitos autorais, especialmente quando técnicas de IA podem recombina conteúdos e dificultar rastreabilidade.

Considerações finais

Este artigo partiu de uma pergunta: como produzir história digital diante do colonialismo de dados, isto é, diante de infraestruturas que extraem, classificam, armazenam e tornam desigualmente visíveis dados, sujeitos e instituições?

A resposta defendida não consiste em recusar o digital, mas em transformá-lo em objeto de crítica e em campo de intervenção historiográfica. Produzir história digital, hoje, significa interrogar simultaneamente documentações físicas e digitais, relacionando as fontes, os bancos de dados, os metadados, o código, as plataformas, os contratos de nuvem e as condições materiais de processamento que participam da fabricação da evidência.

Procurou-se argumentar que a soberania digital não é um atributo abstrato, mas uma condição infraestrutural e institucional: sem capacidade local, pública ou negociada em termos transparentes de armazenamento, processamento e preservação, a pesquisa histórica se torna dependente de regimes opacos de acesso.

Segundo, bancos de dados são recortes que delimitam o dizível e o visível. Por isso, quando a escrita da história se apoia em dados massivos e abordagens computacionais, o problema da soberania reaparece como problema epistemológico: como preservar categorias, critérios de coleta, ontologias, indexadores e documentação metodológica ao longo do tempo?

Trata-se de um problema próprio da disciplina da História, mas atravessado por campos e disciplinas cada vez mais conectados pela computação. Assim, longe de autorizar um neopositivismo digital, empirismo ou realismo ingênuo, a expansão do repertório algorítmico reforça a necessidade de uma crítica de fontes estendida à infraestrutura, aos metadados, às plataformas e aos regimes de circulação de dados.

Terceiro, a dependência computacional reatualiza dependências mais antigas, agora articuladas ao extrativismo e às assimetrias do capitalismo global. A concentração de capacidade de treinamento e inferência de IA, a opacidade sobre clusters e métricas, e a centralidade geopolítica de data centers e aceleradores mostram que a soberania na era da IA passa por cadeias globais de valor que conectam mineração, energia, logística, trabalho e mercado. Nesse contexto, a história digital tem papel duplo: é afetada por essas assimetrias e, ao mesmo tempo, pode torná-las empiricamente visíveis por meio de diagnósticos, mapeamentos e reutilização crítica de dados.

A história digital não produz soberania por si só. Ela pode, contudo, contribuir para uma agenda soberana quando participa da construção de condições de auditabilidade, preservação e reuso. Isso ocorre quando procedimentos técnicos para manuseio do suporte digital deixam de ser operações invisíveis e se tornam escolhas documentadas, discutíveis e replicáveis, já praticadas em universidades, laboratórios e repositórios públicos. Nesse sentido, ciência aberta, reuso de dados e práticas de código livre/aberto são suportes que permitem rastrear decisões, explicitar recortes, reproduzir resultados e sustentar evidências ao longo do tempo.

Referências bibliográficas

- AIROLDI, Massimo (2022). *Machine habitus: Toward a sociology of algorithms*. Cambridge: Polity Press.
- BALAKRISHNAN, Gopal; RIBEIRO, Vera (2020). *Um mapa da questão nacional*. Contraponto Editora.
- BENKLER, Yochai (2002). Coase's Penguin, or, Linux and the nature of the firm. Version 0.4.3, Aug. Disponível em: <https://www.benkler.org/CoasesPenguin.PDF>. Acesso em: 22 fev. 2026.
- BIERSTEKER, Thomas J.; WEBER, Cynthia (orgs.) (1996). *State sovereignty as social construct*. Cambridge: Cambridge University Press.
- BOBROV, Evgeny; BRACCO, Laetitia; DACOS, Marin; FRESSENGEAS, Nicolas; HRYNASZKIEWICZ, Iain; IARKEVA, Anastasiia; PERŠIĆ, Ana; PROUDMAN, Vanessa; ROMARY, Laurent; SABO,

- Rania (2025). The Principles of Open Science Monitoring. Zenodo. DOI: 10.5281/zenodo.18310393. Disponível em: <https://zenodo.org/records/18310393>. Acesso em: 22 fev. 2026.
- BODIN, Jean (1576). *Les six livres de la Republique*. Paris: Jacques du Puys.
- BONALDO, Rodrigo Bragio (2024). Três vivas ao naufrágio: sobre a ascensão e queda das humanidades digitais. In: SESQUIM, Ilda Renata Andreatta *et al.* (orgs.). *Hoje, Teoria da História*. Belo Horizonte: Fino Traço, pp. 139-152.
- BULL, Hedley (1977). *The anarchical society: a study of order in world politics*. London: Macmillan.
- BUZAN, Barry (2004). *From international to world society? English school theory and the social structure of globalisation*. Cambridge: Cambridge University Press.
- CASSINO, João Francisco; SOUZA, Joyce; SILVEIRA, Sérgio Amadeu da (orgs.) (2022). *Colonialismo de dados e modulação algorítmica: tecnopolítica, sujeição e guerra neoliberal*. São Paulo: Autonomia Literária.
- COHEN, Daniel J.; ROSENZWEIG, Roy (2006). *Digital history: a guide to gathering, preserving, and presenting the past on the Web*. Philadelphia: University of Pennsylvania Press.
- COMISSÃO EUROPEIA (2024). AI Act enters into force (notícia). 1 ago. Disponível em: https://commission.europa.eu/news-and-media/news/ai-act-enters-force-2024-08-01_en. Acesso em: 20 jan. 2026.
- COTA, Wesley *et al.* (2019). Quantifying echo chamber effects in information spreading over political communication networks. *EPJ Data Science*, v. 8, art. 35. DOI: 10.1140/epjds/s13688-019-0213-9. Disponível em: <https://doi.org/10.1140/epjds/s13688-019-0213-9>. Acesso em: 20 jul. 2025.
- DE BOLLA, Peter (ed.) (2023). *Explorations in the Digital History of Ideas: New Methods and Computational Approaches*. Cambridge: Cambridge University Press. E-book. DOI: 10.1017/9781009263610.
- COULDRY, Nick; MEJIAS, Ulises A. (2019). *The Costs of Connection: How Data Is Colonizing Human Life and Appropriating It for Capitalism*. Stanford: Stanford University Press.
- FOUCAULT, Michel (2005). *Em defesa da sociedade: curso no Collège de France (1975-1976)*. São Paulo: Martins Fontes.
- GIL, Tiago Luís (2024). Sobre big data e neopositivismo digital na pesquisa em história. *Almanack*, [Guarulhos], n. 36, ep00124. DOI: 10.1590/2236-463336ep00124. Disponível em: <https://www.scielo.br/j/alm/a/vbZwHsRnVdLBD7PWW9Qcjt/?format=pdf&lang=pt>. Acesso em: 20 jan. 2026.
- GRUZINSKI, Serge (2012). *Que horas são... lá, no outro lado*. América e islã no limiar da época moderna. Belo Horizonte: Autêntica Editora.
- HAO, Karen (2025). *Empire of AI: Dreams and Nightmares in Sam Altman's OpenAI*. New York: Penguin Press.
- HARDIN, Garrett (1968). The tragedy of the commons. *Science, Washington*, v. 162, n. 3859, p. 1243-1248. DOI: 10.1126/science.162.3859.1243. Disponível em: <https://math.uchicago.edu/~shmuel/Modeling/Hardin%20Tragedy%20of%20the%20Commons.pdf>. Acesso em: 22 fev. 2026.
- HAWKINS, Zoe Jay; LEHDONVIRTA, Vili; WU, Boxi (2025). *AI Compute Sovereignty: Infrastructure Control Across Territories, Cloud Providers, and Accelerators*. Rochester, NY: Social Science

- Research Network, 20 jun. DOI: 10.2139/ssrn.5312977. Disponível em: <<https://ssrn.com/abstract=5312977>>. Acesso em: 20 jan. 2026.
- JARDELINO, Fábio; CAVALCANTI, Davi Barboza; TONIOLO, Bianca Persici (2020). A proliferação das fake news nas eleições brasileiras de 2018. *Comunicação Pública* [online], v. 15, n. 28. DOI: 10.4000/cp.7438. Disponível em: <<http://journals.openedition.org/cp/7438>>. Acesso em: 14 abr. 2023.
- JOCKERS, Matthew L. (2013). *Macroanalysis: Digital Methods and Literary History*. Urbana, IL: University of Illinois Press.
- KANSTEINER, Wulf (2022). Digital doping for historians: Can history, memory, and historical theory be rendered artificially intelligent? *History and Theory*, v. 61, n. 4, pp. 119-133.
- KITTLER, Friedrich A. (1992). There is no software. *Stanford Literature Review*, v. 9, n. 1, pp. 81-90.
- KITTLER, Friedrich (2016). *Mídias Ópticas: Curso em Berlim, 1999*. Tradução de: Markus Hediger. Rio de Janeiro: Contraponto.
- KOSELLECK, Reinhart (1992). Uma história dos conceitos: problemas teóricos e práticos. *Revista Estudos Históricos*, v. 5, n. 10, pp. 134-146.
- KRASNER, Stephen D. (1999). *Sovereignty: organized hypocrisy*. Princeton: Princeton University Press.
- KRASNER, Stephen D. (2001). Rethinking the sovereign state model. *Review of International Studies*, v. 27, n. 1, pp. 17-42.
- KUUKKANEN, Jouni-Matti (2015). *Postnarrativist philosophy of historiography*. London: Palgrave Macmillan.
- LAVAL, Christian; DARDOT, Pierre (2017). O comum, um ensaio sobre a revolução no século 21. Tradução: Renan Porto. *Lugar Comum: estudos de mídia, cultura e democracia*, n. 49. Disponível em: <https://revistas.ufjf.br/index.php/lc/article/view/49475> (PDF em: <https://revistas.ufjf.br/index.php/lc/article/download/49475/26997/138856>). Acesso em: 22 fev. 2026.
- LE GOFF, Jacques (1990). Documento/Monumento. In: LE GOFF, Jacques. *História e memória*. Campinas: Editora da Unicamp.
- BONALDO, Rodrigo Bragio (2024). As palavras e os tokens: Projeção vetorial aplicada ao estudo da semântica dos tempos históricos. *Revista de Teoria da História*, Goiânia, v. 27, n. 1, p. 7-50. DOI: 10.5216/rth.v27i1.79370.
- MOURÃO, Mônica; LEPRI, Adil Giovanni (2020). A política narrativa do Movimento Brasil Livre no documentário Não vai ter golpe! O nascimento de uma nação livre (2019). *Doc On-line*, n. 28, pp. 117-139, set. DOI: 10.25768/20.04.02.28.07. Disponível em: <<http://www.doc.ubi.pt>>. Acesso em: 20 jul. 2025.
- MUSSELL, James (2012). Doing and Making: History as Digital Practice. In: WELLER, Toni (ed.). *History in the Digital Age*. Abingdon: Routledge, p. 79-94
- MÜSSEMANN, Hannah; AGUIRRE, Gisela; ARBOLEDA, Arcila; ROTA, Alesson (2024). *Manual para el uso de métodos digitales en proyectos de humanidades: la experiencia del proyecto Gumelab*. Medellín: Universidad de Antioquia. Disponível em: <<https://hdl.handle.net/20.500.12733/19320>>. Acesso em: 14 set. 2024.
- O'NEIL, Cathy (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown.

- OHARA, João Rodolfo Munhoz (2022). “Apocalípticos e integrados”: historiadores, computadores e a pesquisa histórica. In: NICODEMO, Thiago et al. (Orgs.). *Caminhos da história digital no Brasil*. Vitória: Milfontes, 2022.
- OSTROM, Elinor (2009). Beyond markets and states: polycentric governance of complex economic systems. *Prize Lecture*, December 8. Disponível em: https://www.nobelprize.org/uploads/2018/06/ostrom_lecture.pdf. Acesso em: 22 fev. 2026.
- PEREIRA, Lauisa Rauter (2010). O Conceito de Soberania: dilemas e conflitos na construção e crise do Estado Imperial Brasileiro (1750-1870). *Intellèctus*, v. 9, n. 2.
- PEREIRA, Mateus Henrique de Faria; MARQUES, Mayra; ARAÚJO, Valdei Lopes de; RAMALHO, Walderez (2022). *Bolsotrump: realidades paralelas (2020-2022)*. Rio de Janeiro: Editora FGV.
- RAMON ROTA, A. (2026). Urdiduras para a preservação digital: o uso de abordagens computacionais para construção de LLM arquivista. *Acervo*, Rio de Janeiro, v. 39, n. 1, p. 1-30. DOI: 10.64729/an.acervo.v39i1.2771. Disponível em: <https://revistaacervo.an.gov.br/index.php/revistaacervo/article/view/2771>. Acesso em: 1 jul. 2026.
- ROBERTSON, Stephen (2022). The properties of digital history. *History and Theory*, v. 61, n. 4, p. 86-106. DOI: 10.1111/hith.12286.
- ROSENZWEIG, Roy (2011). *Clio wired: the future of the past in the digital age*. New York: Columbia University Press.
- ROTA, Alesson Ramon (2024). Patrimônias: algoritmo para classificação de patrimônios incômodos (em disputa). Versão 2.0. Campinas: Universidade Estadual de Campinas, Repositório de Dados de Pesquisa da Unicamp. Disponível em: <<https://doi.org/10.25824/redu/35UZZ3>>. Acesso em: 30 jan. 2026.
- ROTA, Alesson Ramon (2025). *Apologia da história nas temporalidades digitais: redes intelectuais, americanismos e disciplina*. Tese (Doutorado em História). Universidade Estadual de Campinas, Campinas.
- ROTA, Alesson Ramon et al. (2023). Uma década conectada: patrimônios incômodos no paradigma 2013-2023. Dataset. Versão 2.0. Campinas: Universidade Estadual de Campinas, Repositório de Dados de Pesquisa da Unicamp. Disponível em: <<https://doi.org/10.25824/redu/IAQJRJ>>. Acesso em: 30 jan. 2026.
- ROTA, Alesson Ramon, Thiago Lima Nicodemo (2023). Arquivos pessoais e redes sociais: o Twitter construído como documento histórico. *Estudos Históricos (Rio de Janeiro)* 36, p. 44-67.
- ROTA, Alesson Ramon; NICODEMO, Thiago Lima (2024). Uma década conectada: patrimônio incômodos no paradigma 2013-2023. In: SILVA, Evander Ruthieri S. da (org.). *Histórias Conectadas no Sul Global*. São Paulo: Editora Todas as Musas.
- SANTAELLA, Lucia (2024). Diagnóstico do contemporâneo. *Estudos Avançados*, São Paulo, v. 38, n. 110, pp. 7-18, jan./abr. DOI: 10.1590/S0103-4014.2024.38110.002.
- SCIELO. SciELO: Scientific Electronic Library Online. Disponível em: <<https://www.scielo.org/>>. Acesso em: 22 fev. 2026.
- SENADO FEDERAL (2023). Projeto de Lei n. 2.338, de 2023: “Dispõe sobre o desenvolvimento, o fomento e o uso ético e responsável da inteligência artificial com base na centralidade da pessoa humana”. Página da matéria (tramitação). Brasília: Senado Federal. Disponível em:

- <<https://www25.senado.leg.br/web/atividade/materias/-/materia/157233>>. Acesso em: 20 jan. 2026.
- SILVA, Ergon Cugler de Moraes; ROCHA, Isabela; VAZ, José Carlos; VENEZIANI, Julia Ribeiro de Almeida; MODANEZ, Camila de Camargo (2025). Contratos, Códigos e Controle: A Influência das Big Techs no Estado Brasileiro. São Paulo, jul. 2025. Disponível em: <<https://projetobrasil.jornalggn.com.br/artigos/documento/contratos-codigos-e-controle-a-influencia-das-big-techs-no-estado-brasileiro/>>. Acesso em: 20 jan. 2026.
- SILVEIRA, Pedro Telles da (2018). *História, técnica e novas mídias*: reflexões sobre a história na era digital. 375 f. Tese (Doutorado em História) – Universidade Federal do Rio Grande do Sul, Porto Alegre. Disponível em: <<https://lume.ufrgs.br/handle/10183/189249>>. Acesso em: 29 set. 2025.
- SUBRAHMANYAM, Sanjay (2023). Why Connected Histories?: some Reflections. *Storia della storiografia: rivista internazionale*, Pisa; Roma, v. 82/83, n. 1-2, pp. 53-81. DOI: 10.19272/202311502003.
- TORGERSON, Jesse W. (2022). Historical practice in the era of digital history. *History and Theory*, v. 61, n. 4, pp. 37-63, 2022.
- UNITED NATIONS (2024). Digital Economy Report 2024: Shaping an Environmentally Sustainable and Inclusive Digital Future. Nova York: United Nations, 2024. Disponível em: <https://unctad.org/system/files/official-document/der2024_en.pdf>. Acesso em: 20 jan. 2026.
- VAN DIJCK, José (2014). Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology. *Surveillance & Society*, v. 12, n. 2, pp. 197-208.

Declaração de Uso de Inteligência Artificial Generativa

O autor do artigo acima identificado declara que ao longo da elaboração deste manuscrito foram utilizadas as seguintes ferramentas de inteligência artificial: ChatGPT/OpenAI; Qwen da Alibaba Cloud, Codex/OpenAI; DeepL; LanguageTool; PyTorch; Python e bibliotecas associadas, com a finalidade de: revisão de linguagem — aprimoramento do estilo textual e da clareza; tradução — processo de conversão do artigo, total ou parcial, para outro idioma; visualização de dados ou imagens — apoio na produção e/ou aprimoramento de gráficos, figuras ou ilustrações; e aprimoramento de códigos para agenciar os itens anteriores de IA. Após o uso das ferramentas, o autor revisou e editou criticamente todo o conteúdo, assumindo total e integral responsabilidade pelo manuscrito publicado, inclusive por eventuais imprecisões ou incorreções.