pyHDB - ferramenta heurística para a Hemeroteca Digital Brasileira: utilizando técnicas de *web scraping* para a pesquisa em história

pyHDB - heuristic tool for the Brazilian Newspaper Digital Library: using web scraping technics for Historical research

Eric Brasil [a]

ericbrasiln@proton.me

https://orcid.org/0000-0001-5067-8475

[a] Universidade da Integração Internacional da Lusofonia Afro-brasileira, Instituto de Humanidades e Letras, São Francisco do Conde, BA, Brasil.

Resumo

Este artigo tem como objetivo analisar a relação entre ferramentas e interfaces de busca em repositórios de fontes digitais e a construção do conhecimento histórico na era digital. Para tanto, é analisada a *pyHDB: ferramenta heurística para a Hemeroteca Digital Brasileira* da Biblioteca Nacional, caracterizando-se seus aspectos técnicos, metodológicos e heurísticos. Tal ferramenta é um programa de computador escrito com a linguagem de programação *Python* e que utiliza técnicas de *web scraping*. Foi criada com o objetivo de auxiliar pesquisadores no processo de construção e registro metodológico, gerando relatórios e fornecendo dados tabulares e *data sets* a partir dos parâmetros de busca definidos. Primeiramente, são analisados de forma crítica os resultados produzidos pela interface gráfica da Hemeroteca Digital Brasileira. Em seguida, a *pyHDB* é apresentada detalhadamente, tanto em seus aspectos éticos e técnicos quanto em termos de possibilidades analíticas, por meio de três exemplos de busca. Por fim, nas considerações finais, discorre-se acerca das vantagens do desenvolvimento e uso de ferramentas metodológicas digitais para a pesquisa histórica.

Palavras-chave

Metodologia da história; Heurística; História Digital.

Abstract

This article aims to analyze the relationship between search tools and users' interfaces in digital source repositories and the construction of historical knowledge in the digital age. Therefore, I analyze the pyHDB: Heuristic Tool for the Brazilian Digital Newspaper Library of the National Library, characterizing its technical, methodological and heuristic aspects. The tool is a computer program written in the Python programming language and uses web scraping techniques. Its purpose is to assist researchers in the process of methodological construction and recording, creating reports, tabular data and datasets from the defined search parameters. First, the results generated by the Hemeroteca Digital Brasileira graphical interface are critically analyzed. Then, the *pyHDB,* both its ethical and technical aspects and analytical possibilities, is presented in detail through three search examples. Finally, in the concluding remarks, the advantages of developing and using digital methodological tools for historical research are discussed.

Keywords

History Methodology; Heuristics; Digital History.

Entre os dias 11 e 26 de abril de 2021, todas as páginas da Fundação Biblioteca Nacional ficaram fora do ar. Na conta oficial da instituição no Twitter anunciava-se, no dia 12 de abril, que "Por motivos técnicos, nosso site se encontra temporariamente fora do ar. Esperamos solucionar o problema o quanto antes" (FUNDAÇÃO BIBLIOTECA NACIONAL, 2021). A preocupação, em muitos casos, foi visível nas redes sociais, especialmente diante das incertezas e demora na divulgação de informações atualizadas. Pesquisadoras e pesquisadores, de todos os níveis, variadas formações e áreas, queixavam-se em posts e tuítes, e alguns, inclusive, indignados, expressavam-se em caixa alta: "PRECISO TERMINAR MEU TCC".

Este caso aponta para uma questão bem mais ampla, a de uma profunda virada digital nas humanidades nos últimos 20 anos (BRESCIANO, 2000, p. 7-9), e, também, para as modificações ocorridas na História como disciplina. Para a pesquisa histórica, esta mudança reside na combinação entre a digitalização das fontes primárias, a criação profusa e disponibilização de novas fontes nativamente digitais e a *datificação* das relações sociais (SOUTHERTON, 2020) e a consequente transformação nos métodos de pesquisa e construção do conhecimento histórico. Nas palavras de Bob Nicholson, na década de 2010, nós estaríamos "à beira da virada digital na pesquisa em humanidades impulsionada pelo uso criativo dos arquivos on-line e um desejo de imaginar novos tipos de pesquisa" (NICHOLSON, 2013, p. 63).

Se no início da terceira década do século XXI ainda não encontramos um campo consolidado para esses debates na historiografia brasileira – apesar de importantes trabalhos publicados, como os de Anita Lucchesi (2014), Pedro Telles (2018), Thiago Nicodemo (2019), Marcela Albaine Costa (2021) para citar apenas alguns, e do surgimento de laboratórios e investigações sobre o tema –, percebemos que tais questões foram aceleradas e aprofundadas, mesmo que de forma açodada e desequilibrada, pela pandemia da COVID-19 que assola a humanidade desde o início de 2020.

Muitas reflexões emanam desse cenário, ainda que de forma incipiente, em espaços acadêmicos nacionais, girando, em grande medida, em torno de dois polos: de um lado, a urgência em enfrentar tal virada digital como elemento chave para a produção historiográfica do século XXI (ROMEIN *et al.*, 2020, p. 293); e, de outro, a setorização dos aspectos digitais da disciplina em uma subárea, restrita apenas aos interessados "nessas novidades tecnológicas".

Este debate tem muito a crescer a partir do momento em que historiadores e historiadoras das mais distintas filiações teóricas o encararem, dispondo-se a colaborar ativamente na construção de reflexões junto a arquivistas, bibliotecários, programadores, cientistas de dados. Em outras palavras, isso significa atuar no que Max Kemman (2021) chama de "zonas de troca", buscando ampliar tanto os aspectos técnicos quanto epistemológicos do desenvolvimento e uso de ferramentas e métodos digitais para a pesquisa em história.

Neste artigo, enfrento o problema comum a todos e todas que executam a operação historiográfica: o processo de selecionar, recolher, organizar as fontes primárias e realizar a crítica relativa a elas – a chamada heurística das fontes. Agora, assumindo um aspecto diferente, trata-se de uma heurística em ambientes digitais, mediada por de ferramentas e dados digitais. Para tanto, apresento e caracterizo uma ferramenta de suporte para a pesquisa na Hemeroteca Digital Brasileira, da Biblioteca Nacional, intitulada *pyHDB: ferramenta heurística para a Hemeroteca Digital Brasileira,* e analiso sua relação com a teoria, a metodologia e a epistemologia da história. A ideia é aproximar os saberes fundamentais da pesquisa em história com conhecimentos técnicos de programação, atuando nessas zonas de troca, interdisciplinares e colaborativas. Ou, ainda, parafraseando a Fickers e Clavert, busco "combinar habilidades digitais críticas com uma abordagem autorreflexiva", prática chamada de hermenêutica digital. Isto implicaria "[t]ornar explícito como a produção de conhecimento histórico por meio de ferramentas e tecnologias digitais é o resultado de um processo complexo de interação humano-máquina, de coconstrução do 'objeto epistêmico' da inquirição e investigação histórica" (CLAVERT; FICKERS;, 2021, n.p., tradução nossa).

Argumento que as novas formas de realizar pesquisas em interfaces digitais de busca impactam, mediam e direcionam tanto a coleta e seleção das fontes quanto sua análise. Diante disso, é fundamental que o método histórico leve em consideração a aplicação de práticas de heurística digital coerentes tanto com as características das ferramentas e métodos utilizados, das fontes e dados trabalhados quanto com as reflexões teóricas básicas da disciplina histórica.

A ferramenta heurística para a Hemeroteca Digital Brasileira (*pyHDB*) é um programa de computador escrito com a linguagem de programação Python e desenvolvido por mim no âmbito das pesquisas que realizo no curso de História do Instituto de Humanidades e Letras do campus dos Malês da Universidade da Integração Internacional da Lusofonia Afro-brasileira (UNILAB) e no Laboratório de Humanidades Digitais da Universidade Federal da Bahia (UFBA). A ferramenta tem como principal objetivo auxiliar pesquisadores no processo de construção e registro de dados com rigor metodológico por meio de relatórios, planilhas e *data sets* (conjuntos de dados) criados a partir dos parâmetros de busca disponibilizados pela interface da Hemeroteca Digital Brasileira.[1] A *pyHDB* registra tanto os parâmetros definidos pelo pesquisador quanto a hora e data da busca, os resultados encontrados, os correlativos acervos, o número de ocorrências, o número de páginas de cada acervo buscado, a frequência de ocorrências por páginas e as informações de cada ocorrência e realiza o download da imagem quando permitido pela Hemeroteca Digital Brasileira.

[1] Ao longo do artigo, a interface de busca, acessível pelo link http://memoria.bn.br/, será referida como Hemeroteca Digital Brasileira, enquanto a ferramenta metodológica desenvolvida por mim, será referida como *pyHDB*.

Nas páginas que se seguem, primeiramente, analiso de forma crítica os resultados originados pela interface gráfica da Hemeroteca Digital Brasileira. Em seguida, apresento detalhadamente a pyHDB, tanto em seus aspectos éticos e técnicos quanto em termos de possibilidades analíticas, por meio de três exemplos de busca. Por fim, proponho algumas considerações finais acerca de: a) as vantagens do desenvolvimento de ferramentas metodológicas digitais específicas de acordo com interesses e perguntas de pesquisa e b) as vantagens para a Hemeroteca Digital Brasileira e para a pesquisa histórica em geral da sofisticação das ferramentas disponíveis para o acesso dos usuários e o engajamento de historiadores/as nesse processo.

A interface da Hemeroteca Digital Brasileira: busca, resultados e impactos na pesquisa

Muitas ferramentas e métodos digitais são rapidamente repelidos em contextos acadêmicos, taxados como inviáveis seja por sua curva de aprendizado muito grande ou por sua opacidade técnica (NASCIMENTO, 2020). Entretanto, seus usos são cada vez mais comuns e mesmo imprescindíveis para a execução de pesquisas nas ciências humanas, abrangendo desde o uso do e-mail institucional até algoritmos que processam linguagem natural, passando por editores de textos e planilhas, bancos de dados, mapas, etc. De fato, os programas de computador, que nada mais são do que uma longa sequência de códigos, caminham lado a lado com nossas pesquisas, aulas e publicações.

Contudo, como afirmam Röhle e Rieder, mesmo que uma parte significativa de nossas horas de trabalho transcorra na frente de uma tela, muitas das ferramentas que utilizamos não são enquadradas como parte de nossos métodos de pesquisa. Os autores entendem que a "função heurística dos métodos digitais de pesquisa" estaria focada em encontrar padrões, dinâmicas e relações nos dados e que, mesmo que não estejam muito visíveis na cadeia metodológica, teriam repercussões epistemológicas significativas: computadores como ferramentas heurísticas "oferecem-nos perspectivas particulares sobre os fenômenos nos quais estamos interessados" (RÖHLE; RIEDER, 2012, p. 70, tradução nossa).

No caso específico da Hemeroteca Digital Brasileira e sua interface de busca, é fácil constatar seu valor, importância e impacto positivo na pesquisa e imaginação histórica, como já venho apontando em pesquisas recentes (BRASIL; NASCIMENTO, 2020). Seu uso tem se popularizado em pesquisas acadêmicas, e ela pode ser considerada o principal caminho de acesso às fontes de imprensa no Brasil, sobretudo no que diz respeito a jornais e revistas

do século XIX até meados do século XX. Como veremos detalhadamente, essa interface de busca reúne os predicados de uma ferramenta heurística que, ao mesmo tempo, apresenta potencialidades enormes para a pesquisa em história e impacta diretamente nas formas de construção do conhecimento. Nesse sentido, proponho a seguir uma análise da interface e das possibilidades de resultados.

A Hemeroteca Digital Brasileira foi lançada em 2012 com aproximadamente 5 milhões de páginas digitalizadas disponíveis para consulta e busca por palavras-chave (BETTENCOURT; PINTO, 2013). Atualmente, esse número ultrapassa os 20 milhões de páginas. É possível acessar o acervo de diversas formas: lendo os jornais de forma corrente, filtrando datas, locais, jornais ou – e este é o grande diferencial oferecido pela ferramenta – buscando por palavras-chave combinadas a diferentes filtros, o que proporciona resultados bastante específicos para cada conjunto de parâmetros de busca (BRASIL; NASCIMENTO, 2020).

Assim como em qualquer interface do usuário (UI) – termo traduzido do inglês *user interface*, designando um conjunto de controles por meio dos quais um usuário humano interage com uma máquina (RAYMOND, 2003, p. 289) –, o resultado retornado para o usuário da Hemeroteca Digital Brasileira apresenta uma série de características que direcionam uma determinada maneira de lidar com os dados obtidos e, consequentemente, informam construções metodológicas e epistemológicas moldadas por esse próprio modelo de resultado.

Um jornal digitalizado, nas palavras de Maud Ehrmann e colaboradores, "é um objeto complexo determinado por múltiplas camadas de processamento e dataficação" (EHRMANN; BUNOUT; DÜRING, 2019, p. 2, tradução nossa). As interfaces dos repositórios digitais, portanto, têm um papel central na relação do usuário com a fonte: "Não apenas elas controlam o que os usuários podem aprender sobre o conteúdo digitalizado; elas também moldam ativamente os fluxos de trabalho do usuário, oferecendo diferentes seleções de ferramentas e recursos para pesquisar e explorar esse conteúdo" (EHRMANN; BUNOUT; DÜRING, 2019, p. 2, tradução nossa). Ao mesmo tempo, comumente os usuários dessas interfaces "não estão conscientes dos vieses nos resultados de buscas, causados pelo processamento e dataficação dos jornais" (PFANZELTER *et al.*, 2020, p. 1, tradução nossa).

Porém, é importante lembrar que esses pontos se aplicam a qualquer arquivo digital. Como argumenta Helle Jensen,

> A literacia arquivística digital requer a compreensão de como a produção de arquivos digitais se baseia em designs técnicos que influenciam a sua usabilidade. Isso

> significa que (todos) os historiadores precisam adquirir competência digital em um nível profissional paralelo ao das habilidades que temos para compreender como a classificação e categorização de fontes afetam nossa interação com arquivos analógicos e moldam nossas questões de pesquisa (JENSEN, 2021, p. 6, tradução nossa).

A autora elenca três pontos fundamentais que devem ser encarados criticamente por historiadores/as ao lidarem com arquivos digitais: a) categorias e etiquetas predefinidas; b) pesquisa por campos e organização de resultados; c) metadados. As categorias e etiquetas estruturam o material arquivado de forma a "codificar" leituras e usos específicos; as pesquisas por campo e organização de resultados sugerem certos usos de seus recursos, mesmo quando os usuários não precisam seguir esses "usos preferenciais"; e os metadados – dados sobre os dados –, que muitas vezes são o principal caminho para realizar buscas nos acervos e, portanto, seu registro – parcial ou incompleto –, imputam certas interpretações às fontes (JENSEN, 2021, p. 7-8).

Johan Jarlbrink e Pelle Snickars nos lembram algo que deveríamos levar sempre em consideração ao clicarmos no botão de busca de qualquer acervo digital: que a "digitalização de um jornal histórico não é um processo neutro onde os dados são transferidos de um meio a outro. Ao contrário, quando jornais são digitalizados, eles são transformados." Portanto, como também apontam os autores, as maneiras como os dados digitalizados (como arquivos de jornais) são "criados, armazenados, processados e formatados naturalmente têm implicações em como os registros históricos, como jornais digitais, podem ser acessados e usados, bem como nas histórias que poderão ser exploradas e nas histórias que poderão ser (re)contadas." (JARLBRINK; SNICKARS, 2017, p. 1229–1230, tradução nossa).

Voltemos à página com os resultados retornados após uma busca na Hemeroteca Digital Brasileira. Em qualquer uma das categorias de busca (*Período, Periódico, Local*), os resultados são apresentados de forma individualizada organizada por acervos.[2] Se o jornal possui apenas um acervo, o resultado retornado é a página com a primeira ocorrência e o acesso se dá de ocorrência em ocorrência, sem a possibilidade, por exemplo, de uma percepção estruturada da distribuição das ocorrências no tempo. Em periódicos que possuem vários acervos, é retornada uma lista com os acervos que contiverem ocorrências da busca, sendo classificada pelo número de ocorrências decrescentes.

[2] Termo utilizado pela Hemeroteca Digital Brasileira para designar as pastas de cada periódico, organizadas, geralmente, por recortes temporais de uma década (ex.: 1900-1909).

Assim, os resultados direcionam o acesso às ocorrências numa ordem específica – pautada pelo padrão do próprio site. Mesmo que o usuário classifique as ocorrências por título de acervos, muitas vezes teremos uma lista com sobreposição de períodos, visto que muitos periódicos têm acervos organizados em outros padrões diferentes daquele da estrutura de décadas que pauta a ferramenta de busca.

Ao acessar um desses acervos para ler as ocorrências, o usuário será guiado pela estrutura e pela própria concepção técnica da ferramenta a seguir as ocorrências, mas não terá acesso organizado ao conjunto das ocorrências no interior do acervo e em relação ao todo dos resultados da busca.

Essas características favorecem uma heurística de "proximidade virtual", conceito definido por Janine Solberg (2012) como o potencial de encontrar fontes on-line e de selecioná-las a partir de uma percepção de "proximidade" possibilitada por tecnologias de organização, busca e recuperação. No caso da Hemeroteca Digital Brasileira, seu sistema de busca simples e o retorno de resultados abrangendo décadas e inúmeros jornais diferentes produzem nos usuários a sensação de "proximidade" entre os resultados e reforçam o uso tanto da ferramenta quanto das fontes por ela selecionadas.

Além disso, nesse padrão de resultado, o usuário não tem a possibilidade de acessar o arquivo de texto com o reconhecimento ótico de caracteres – OCR – sobre o qual a busca foi realizada, nem há dados sobre a precisão do OCR para cada resultado específico. Mesmo com as melhorias significativas dos últimos anos, quanto mais antigo o jornal, mais difícil resulta o reconhecimento ótico de caracteres, o que é agravado pela qualidade do original ou do microfilme que serviu de suporte para a digitalização, pelo tamanho reduzido de algumas fontes e pelos seus tipos variados (SALMI, 2020, p. 47).

O OCR realizado no acervo da Hemeroteca Digital Brasileira utilizou o software *ABBY FineReader 11 Professional*. Para a indexação das palavras do conteúdo foi utilizada a tecnologia *DocPro*, chamada "Inteligenciamento DocPro", cujo objetivo é otimizar os resultados ao basear as buscas por aproximação visual (BETTENCOURT; PINTO, 2013, p. 7). Aqueles que utilizam diversos tipos de busca em OCRs pela internet percebem a qualidade elevada e a boa precisão das buscas realizadas na Hemeroteca Digital Brasileira. Entretanto, os dados desse processo, o algoritmo e metadados que o tornam possível, não são disponibilizados para o usuário.

Segundo Angela Maria Bettencourt e Monica Pinto, a indexação dos metadados dos acervos digitalizados utilizou o padrão da BNDigital (autoria, título, assuntos, datas e coleção) e também o padrão *Dublin Core* (BETTENCOURT; PINTO, 2013, p. 5). Este padrão de criação de metadados, lançado em 1995, possui atualmente 15 elementos e é bastante completo (WEIBEL,

1995). Entretanto, ele não está disponível para os usuários. Apesar de o Dublin Core utilizar a linguagem de marcação XML – *Extensible Markup Language* –, não há arquivos XML das páginas digitalizadas disponíveis para o usuário.

Dessa maneira, o procedimento heurístico, a execução da crítica à fonte, fica restrito à emulação de procedimentos analógicos diante de arquivos digitais. Replico aqui a pergunta de Jarlbrink e Snickars: "Como seria possível praticar a crítica de fontes quando os mecanismos e algoritmos para selecionar, capturar, processar e armazenar os dados históricos estão escondidos atrás de interfaces gráficas de usuário?" (JARLBRINK; SNICKARS, 2017, p. 1229, tradução nossa).

Para Salmi, o uso dessas interfaces gráficas é essencial para os historiadores. Entretanto,

> para a posterior análise digital do material, é necessário ser capaz de fazer o download dos dados, por exemplo, como um dump de dados que inclua os textos OCR e metadados, muitas vezes em padrões XML internacionais, como os formatos METS e ALTO (SALMI, 2020, p. 50, tradução nossa).

Portanto, entendo que os pesquisadores e pesquisadoras que utilizam a interface da Hemeroteca Digital Brasileira têm acesso a apenas um pequeno fragmento das possibilidades abertas pelo volume enorme de dados que compõem o acervo. O *corpus* de milhões de páginas com OCR não pode ser acessado como *Big Data*, e os usos de técnicas de leitura distante – *distant reading* (GOODING; TERRAS; WARWICK, 2013) – e do Processamento de Linguagem Natural (PIROVANI; OLIVEIRA, 2018) se mantêm fora de alcance no momento.

Outra questão importante a considerar é a aplicação, na interface da Hemeroteca Digital Brasileira, dos parâmetros de pesquisa e sua relação com os resultados. Nenhum dos parâmetros utilizados na busca é retornado na página de resultados, além de o termo da busca aparecer em um box no topo da página. Desta maneira, não há registro das opções iniciais utilizadas para a busca: se o parâmetro inicial foi "período", "periódico" ou "local" e, mais importante, quais as combinações entre essas variáveis? Esses parâmetros também não são registrados no hiperlink, a URL (*Uniform Resource Locator*) da página específica de uma dada ocorrência; nem a data ou horário da busca, o que seria fundamental para localizar a pesquisa no tempo e, dessa forma, contar com um indicativo do momento em que o acervo da Hemeroteca Digital Brasileira contava com aquele conjunto de páginas digitalizadas e disponibilizadas.

O registro desses procedimentos metodológicos é fundamental para a pesquisa, visto que a menor alteração em uma dessas múltiplas opções geraria resultados distintos (BRASIL;

NASCIMENTO, 2020, p. 213). A própria inclusão ou não do termo ou termos de busca entre aspas duplas altera completamente o resultado, pois define se a busca deve ser realizada com uma expressão exata ou não.

Assim, o motor de busca combinado com as decisões técnicas sobre como apresentar os resultados e quais elementos o usuário pode recuperar e baixar estreitam as possibilidades de uso desse arquivo monumental e fundamental para a pesquisa histórica. Mesmo sendo possível encontrar uma palavra ou frase entre milhões de páginas – o que é um recurso fenomenal que tem transformado a pesquisa no Brasil –, não é possível trabalhar com esses mesmos resultados servindo-se de métodos e ferramentas mais sofisticadas das humanidades digitais. Nesse sentido, evidenciamos como as interfaces não são veículos transparentes, mas sim parte de uma cadeia de operações mediadas (GALLOWAY, 2012).

Todas essas características, entretanto, não invalidam nem devem desmotivar o uso da Hemeroteca Digital Brasileira. Pelo contrário, a própria ferramenta oferece recursos para a superação de muitas questões relacionadas à crítica das fontes (acesso às edições completas, acesso a todas as edições, acesso a informações sobre cada acervo, etc.), aspecto que discuti com profundidade em pesquisa anterior realizada junto com Leonardo Nascimento (BRASIL; NASCIMENTO, 2020). Além disso, a constância e integridade da interface favorecem sua utilização, garantindo aos usuários condições de conhecer sua estrutura e repetir suas buscas ao longo do tempo.

Também é importante destacar o papel fundamental da Fundação Biblioteca Nacional para a preservação, divulgação e pesquisa da história no Brasil. Todo o trabalho, muitas vezes hercúleo, de seus funcionários e funcionárias em prol da manutenção e expansão dos acervos e da própria instituição em meio a políticas de constante ataque à ciência e redução de investimentos no setor nos últimos anos merece não apenas destaque, mas também apoio, empatia e colaboração de toda sociedade brasileira.

Codificando como um historiador

Os elementos apresentados até aqui, que podem ser resumidos na opacidade dos algoritmos e características da interface de busca, exemplificam bem aquilo para o qual Mats Fridlund e colaboradores chamaram atenção em obra recente:

> Os métodos de pesquisa digital criam demandas novas e às vezes mais rigorosas de precisão, pensamento metodológico, auto-organização e colaboração do

> que a pesquisa histórica tradicional (FRIDLUND; OIVA; PAJU, 2020, p. 14, tradução nossa).

Buscando contribuir para a elaboração de caminhos heurísticos digitalmente críticos e conscientes, e entendendo que os métodos digitais de pesquisa, cada vez mais recorrentes, demandam de nós esse cuidado, apresentarei e analisarei a *pyHDB,* ferramenta criada a partir de interesses específicos da pesquisa de um historiador utilizando a Hemeroteca Digital Brasileira.

O objetivo da ferramenta é auxiliar metodologicamente pesquisas realizadas no acervo da Hemeroteca Digital Brasileira, garantindo recursos para uma heurística digital, e, ao mesmo tempo, potencializar o tratamento e análise das fontes digitalizadas, garantindo a transparência metodológica no momento da escrita e apresentação dos resultados.

Assim, a *pyHDB* realiza a coleta automatizada de dados a partir dos parâmetros definidos pelo usuário junto à interface da Hemeroteca Digital Brasileira. A técnica de coletar dados da *Web* de forma automatizada é chamada de *web scraping*. Segundo Noortje Marres e Esther Weltevrede, podemos definir *scraping* da seguinte maneira:

> A raspagem, dito de maneira bastante formal, é uma técnica importante para a coleta automatizada de dados on-line. É uma das práticas mais distintas associadas às formas atuais de pesquisa social digital, aquelas que são marcadas pelo surgimento da Internet e a nova onipresença dos dados digitais na vida social. *Scrapers*, para dizer de forma mais informal, são bits de código de software que possibilitam o download automático de dados da *Web* e a captura de algumas das grandes quantidades de dados sobre a vida social que estão disponíveis em plataformas on-line como Google, Twitter e Wikipédia (MARRES; WELTEVREDE, 2013, p. 313, tradução nossa).

As autoras, pensando a relação do *web scraping* com a pesquisa na área das ciências sociais, afirmam que tal técnica possui capacidade de transformar a pesquisa social e reconfigurar as relações entre sujeitos, objetos, métodos e técnicas da pesquisa pelo fato de ser uma técnica de coleta de informação estruturada a partir de ambientes de informação digital heterogeneamente ordenados. Ou seja, criar um *scraper* possibilita uma solução para enfrentar o problema de encontrar relevância na abundância de dados digitais, visto que "provê um caminho para extrair campos específicos ou elementos de dados de páginas da *Web* e outras fontes da Internet,

transformando dados on-line em conjuntos de dados usáveis e bem-ordenados" (MARRES; WELTEVREDE, 2013, p. 316).

Segundo Vlad Krotov e colaboradores, o *web scraping* possui três fases interligadas: análise da página *web*; rastreio ou raspagem da página *web*; e organização dos dados.

> Cada fase requer o entendimento de várias tecnologias da *Web* e pelo menos uma linguagem de programação popular, como R ou Python. No entanto, essas três fases geralmente requerem pelo menos algum envolvimento humano e, portanto, não se pode automatizá-las totalmente (KROTOV; JOHNSON; SILVA, 2020, p. 540, tradução nossa).

Programas que se baseiam em técnicas de *web scraping* apresentam também vulnerabilidades e problemas que merecem atenção. A instabilidade é recorrente, pois necessitamos extrair dados específicos de lugares específicos na *Web*. Assim, se o layout e estrutura da página ou do site mudarem, ajustes obrigatórios terão de ser feitos no código-fonte do programa. A curva de aprendizado para a construção de ferramentas próprias é muito grande e demanda conhecimentos técnicos específicos, o que inevitavelmente limita o seu uso.

Entretanto, o *web scraping* não deve ser entendido apenas como uma técnica, mas também como uma maneira particular de lidar com a informação e o conhecimento. Portanto, "também é uma prática de análise" (MARRES; WELTEVREDE, 2013, p. 317, tradução nossa). As escolhas e perguntas, interesses e abordagens do/a pesquisador/a determinam os critérios de coleta que serão implementados pelo programa. Como afirmaram Röhle e Rieder,

> nossos ajudantes digitais já estão cheios de "teoria" e "julgamento". Como acontece com qualquer metodologia, eles contam com conjuntos de premissas, modelos e estratégias. A teoria já está trabalhando no nível mais básico quando se trata de definir unidades de análise, algoritmos e procedimentos de visualização (RÖHLE; RIEDER, 2012, p. 70, tradução nossa).

Logo, as ferramentas digitais são sempre, ao mesmo tempo, metodológicas e profundamente imiscuídas de teoria, o que impacta nos resultados epistemológicos.

Podemos pensar no exemplo de um pesquisador que pretenda raspar uma base de dados de estudantes de uma universidade. Entre os elementos de uma longa série de informações disponíveis no repositório on-line da instituição, aparece o campo "raça/etnia" ao lado de campos relativos a renda, escolaridade, gênero, entre outros. O pesquisador constrói um programa para coletar esses dados de forma automatizada e estruturada, alimentando um banco de dados que será a base de uma pesquisa sobre experiências estudantis. No entanto, ele escolhe não coletar o campo "raça/etnia" por entender que esse elemento não é relevante para sua pesquisa. Ou, ainda, podemos imaginar que, por um erro no código do programa, esse campo não é coletado corretamente. De qualquer modo, em função de decisões teóricas sobre os campos e categorias relevantes para a pesquisa ou de erros técnicos, teremos um *data set* específico e com vieses bem marcados alterando a produção do conhecimento possível a partir dele.

Com efeito, como pesquisadores e pesquisadoras, principalmente negras e latinas, já vêm afirmando e comprovando há bastante tempo, toda tecnologia é repleta de vieses e reproduz as desigualdades e violências sociais e, por isso, deve ser submetida a um forte e constante escrutínio crítico (BENJAMIN, 2019a, 2019b; BIRHANE, 2021; NOBLE, 2018; SILVA, 2020).

Aspectos éticos

Antes de avançar para os critérios que serviram como parâmetros para a construção da ferramenta *pyHDB*, precisamos enfrentar algumas questões éticas que envolvem o uso de programas de *web scraping*. Inspirado na abordagem baseada em princípios proposta por Mathew Salganik (2017, pp. 282-283), todo o processo de elaboração da ferramenta *pyHDB* procurou lidar eticamente com o acervo da Hemeroteca Digital Brasileira. Procurei equilibrar os quatro princípios fundamentais apresentados por Salganik: respeito pelas pessoas; benefícios; justiça e respeito pela lei; e interesse público.

Ao construir o código, intervalos de tempo foram estabelecidos com vistas a não sobrecarregar os servidores da instituição, bem como limites de tempo para aguardar a resposta do servidor; procurei documentar todo o código tornando-o público e humanamente legível. Os dados coletados são apenas aqueles abertos ao público para pesquisa e, quando o periódico está sob alguma restrição legal, conforme informação fornecida pela própria interface da Hemeroteca Digital Brasileira, apenas os dados são coletados, sem o download da imagem.

Essa questão que envolve direitos autorais poderia afetar ou mesmo enviesar os resultados do uso da ferramenta, visto que apenas as imagens dos jornais sem restrições são salvas localmente. Entretanto, os links de todas as páginas dos jornais com restrições de direitos

autorais são registrados no *data set*, assim como a informação de que não consta imagem baixada. Isto permite ao usuário remediar a questão mapeando as páginas inexistentes localmente e acessando-as diretamente na base da Hemeroteca Digital Brasileira com o hyperlink único.

Não constam na página inicial da Biblioteca Nacional Digital (http://bndigital.bn.gov.br/) os termos de uso referentes à pesquisa na interface da Hemeroteca Digital Brasileira. O arquivo robots.txt presente na página da BNDigital (http://bndigital.bn.gov.br/robots.txt) informa que há restrições para o acesso automatizado apenas na página wp-admin/, mas não encontramos nenhum arquivo para o domínio http://memoria.bn.br/ onde se informasse sobre quaisquer limitações de interação entre robôs e a página. O arquivo robots.txt é uma sintaxe chamada Padrão de Exclusão de Robôs criada em 1994. Nele, são apresentadas regras que se aplicam a determinado site em relação à atuação de robôs (MITCHELL, 2018, p. 313).

Ao mesmo tempo, o código, o *data set* usado no artigo e as análises estão disponíveis de forma pública e trazem benefício à comunidade. As ferramentas utilizadas e produzidas têm código aberto e licenças de uso MIT, sem nenhum interesse financeiro. Entendo que os benefícios e interesse público, assim como o respeito pela ciência aberta e sua divulgação e o respeito aos avisos legais informados pela interface garantem o caráter ético desta pesquisa.

Colocando a *pyHDB* em diálogo com outras iniciativas

Outras iniciativas muito relevantes no que se refere à elaboração de ferramentas e métodos digitais para a pesquisa nas humanidades têm sido desenvolvidas recentemente e, mesmo que uma análise aprofundada de suas especificidades fuja do escopo deste artigo, é importante fazer uma breve apresentação delas para localizar cientificamente a *pyHDB*.

O desenvolvimento de ferramentas digitais de coleta, organização e análise de dados (sejam textos, imagens, sons, etc.) demanda muito investimento de tempo e capital, alto poder computacional, equipes multidisciplinares e grande conhecimento técnico, além de teórico. Assim, as iniciativas listadas aqui são resultado de projetos com alto grau de financiamento, que contaram com prazos longos e equipes grandes e variadas.

Muitos desses projetos são a representação do acúmulo de produção de laboratórios inteiros, como o conjunto de 30 ferramentas para coleta, organização, visualização, escrita criadas e disponibilizadas pelo MediaLab da universidade *Sciences Po;* ou são o produto da parceria entre várias iniciativas, como o GLAM Workbench (SHERRATT, 2021), que reúne ferramentas, tutoriais, dicas e exemplos para dar suporte a pesquisas com GLAM (*Galleries, Libraries, Archives & Museums*), principalmente pela disponibilização de *Jupyter Notebooks*.

Especificamente sobre jornais históricos, o projeto *impresso* oferece um aplicativo on-line para pesquisa, análise e visualização de dados e textos minerados de mais de 200 anos de jornais europeus. O projeto, uma parceria do Laboratório de Humanidades Digitais do Instituto Federal Suíço de Tecnologia em Lausanne com o Instituto de Linguística Computacional da Universidade de Zurique e o Centro de História Contemporânea e Digital da Universidade de Luxemburgo, oferece recursos muito sofisticados de pesquisa e análise, com linhas do tempo, *ngrams*, tópicos, além da busca e acesso a metadados, texto gerado por OCR, criação de coleções pelo próprio usuário, recursos de comparação, entre outros.

Importante também é o ambiente *web* de análise e leitura de texto Voyant Tools (SINCLAIR, ROCKWELL, 2016), que possibilita a análise e visualizações variadas de *corpora* documentais diretamente no navegador *web*. Ou, ainda, o livro/curso de Melanie Walsh *Introduction to Cultural Analytics & Python*, onde a autora apresenta uma série de técnicas organizadas em *Jupyter Notebooks* voltadas para a análise de texto – especialmente com as bibliotecas Pandas e Plotly, também utilizadas neste artigo (WALSH, 2021). O conjunto de tutoriais, revisados por pares, publicados pelo periódico acadêmico de metodologia para historiadores digitais *Programming Historian*, em suas versões em inglês (2012), espanhol (2016), francês (2019) e português (2021), tem contribuído de maneira constante para a consolidação do uso crítico e consciente de ferramentas e métodos digitais no campo da história. Abrangendo um conjunto variado de temas, as publicações tratam de todas as etapas da pesquisa e escrita da história, utilizando diferentes linguagens de programação.

Todos os exemplos aqui citados são de código aberto e representam esforços coletivos e interdisciplinares, com financiamentos variados, mas significativos para a realização dos objetivos visados. A *pyHDB* se inspira sobremaneira no caráter aberto, colaborativo e interdisciplinar dos exemplos apresentados e, sobretudo, busca se inserir nas reflexões teóricas sobre seus desenvolvimentos, aplicações e repercussões para o campo da história.

pyHDB: a ferramenta heurística para a Hemeroteca Digital Brasileira

Devido à complexidade da interface da Hemeroteca Digital Brasileira, a construção de uma ferramenta de auxílio metodológico para lidar com a busca e os resultados da pesquisa demanda conhecimento avançado de linguagem de programação e, especificamente, técnicas e bibliotecas mais complexas sobre *web scraping*. Na criação da *pyHDB*, utilizamos a linguagem Python para a escrita do código e um conjunto de módulos e bibliotecas com funcionalidades que atendiam aos objetivos propostos. A ferramenta *pyHDB* é um programa de computador que não

possui interface gráfica de usuário no momento. Portanto, ela é executada a partir do terminal no computador pessoal e, para isso, é preciso ter o Python 3 instalado. O código e a documentação completa estão disponíveis on-line, no repositório do GitHub ericbrasiln/pyHDB, cujo código-fonte também está armazenado em repositório confiável no Zenodo. A versão utilizada neste artigo pode ser acessada pelo DOI 10.5281/zenodo.5696670.

A interface da Hemeroteca Digital Brasileira possui elementos dinâmicos e complexos que se alteram e respondem a interações com o usuário. Portanto, mesmo que a página *web* seja escrita em HTML – *HyperText Markup Language* (TURKEL; CRYMBLE, 2012), o "padrão oculto atrás de quase tudo o que vemos e fazemos ao navegar na *Web*" (NASCIMENTO, 2017, p. 3) –, ela apresenta elementos de *JavaScript* e uso de *caches* que exigiram a utilização da biblioteca *Selenium* para realizar os cliques e ações on-line de forma automatizada. Esta biblioteca possibilita ao computador emular o acesso à página *web* e executar as funções necessárias para nossos interesses (MITCHELL, 2018).

Antes de iniciar a elaboração do código por trás do programa, foi fundamental o acúmulo de conhecimento acerca das possibilidades de uso e do próprio acervo da Hemeroteca Digital Brasileira. Venho utilizando o acervo de jornais e revistas da Biblioteca Nacional Brasileira desde 2009, primeiro com a leitura das edições em microfilme e, desde 2012, pesquisando e produzindo reflexões históricas sobre e a partir da Hemeroteca Digital Brasileira. Mesclando esse arcabouço com meus interesses de pesquisa, pude elaborar os critérios para a construção da ferramenta.

Minhas pesquisas lidam diretamente com as experiências sociais de homens e mulheres negras na cidade do Rio de Janeiro no pós-abolição, debatendo associativismo, cidadania e performances culturais. Preocupado com essa história vista de baixo, tenho buscado acompanhar as trajetórias desses sujeitos na cidade ao longo das primeiras décadas do século XX. Diante disso, metodologicamente é mais coerente e eficiente realizar a busca com o parâmetro inicial de "Local", seguindo metodologia apresentada por Brasil e Nascimento (2020).

Assim, esta versão da *pyHDB* estabelece a seguinte cadeia de parâmetros: 1) Local; 2) Período; 3) Periódico (por padrão, *Todos*); 4) Termo da Busca. Ao iniciar a ferramenta, aparecerá uma breve apresentação na tela de seu computador. Em seguida, é solicitada a definição do local e o recorte temporal. Nesta versão, a busca será efetuada em todos os periódicos existentes para a configuração especificada para "Local" e "Período". Portanto, o parâmetro 3 está definido por padrão como "Todos" no código fonte. O último parâmetro a ser passado pelo usuário é o termo de busca (Figura 1).[3]

[3] Todas as imagens podem ser acessadas em seu tamanho original no link: ericbrasiln.github.io/analise_pyHDB/Imagens.

Figura 1 - Print da tela de execução da ferramenta *pyHDB* mostrando os parâmetros de busca utilizados no exemplo 1 deste artigo

```
=================================================================================

1 - Local
Orientações para busca:
- O termo deve ser idêntico às opções listadas na página da HDB;
- Esse parâmetro é case sensitive;

Digite o local de busca: RJ

2 - Período
Orientações para busca:
- O recorte deve ser escerito de forma idêntica às opções listadas na página da HDB;
- É possível buscar todos os periódicos digitando `Todos`

Digite o período de busca: Todos

3 - Periódico: Todos

4 - Termo da busca
Orientações para busca:
- Coloque o termo entre aspas duplas para expressões exatas;
- Não use acentos;
- Não mais que três palavras

Digite o termo de busca: "germano lopes da silva"
```

Fonte: Elaborado pelo autor a partir de Brasil (2021).

Com os parâmetros de busca definidos pelo usuário, o programa acessa a página inicial da Hemeroteca Digital Brasileira (http://memoria.bn.br/hdb/), insere os parâmetros e clica no botão de pesquisar. Aguarda a página de resultados ser carregada e mostra na tela a quantidade de acervos dos jornais com alguma ocorrência até o limite de 100 jornais (ou seja, a segunda página de resultados, de acordo com o padrão definido, buscando minimizar erros e sobrecargas do servidor).

A partir daí, o programa cria uma listagem com todos os acervos com ocorrências e executa uma iteração nessa lista, acessando cada acervo e coletando as informações detalhadas de cada ocorrência. Se o acervo possuir algum aviso legal de restrição de acesso, este aparecerá na tela e as imagens das páginas com ocorrências não serão baixadas para o computador do usuário.

Devido à característica da interface do usuário e da necessidade de acessar vasto acervo, é comum que aconteçam erros no carregamento de páginas na interação entre usuário e servidor da Hemeroteca Digital Brasileira. Assim, busquei estabelecer limites de tempo para aguardar

a resposta da página, e, se esta não acontece, é gerado um relatório de erros. Ainda assim, podem acontecer erros não antecipados pelo algoritmo da *pyHDB*. Nesses casos, se o programa se encerrar, é possível retomar a partir do último acervo analisado, desde que seja executado novamente com a mesma configuração de parâmetros e na mesma data. A *pyHDB* avalia quais acervos já foram coletados para seu computador e passa para o seguinte.

Também é importante destacar que esses erros são mais comuns conforme o volume de ocorrências aumenta. O programa foi pensado e testado para lidar com até algumas milhares de ocorrências, funcionando bem com resultados com 5 mil ocorrências. Para um funcionamento mais eficiente, é prudente compartimentar as buscas por décadas e realizá-las ao longo da noite, evitando horários com maior utilização da interface.

Resultados gerados pela *pyHDB,*

Após executar a ferramenta *pyHDB*, o usuário recebe um conjunto de resultados que compreende o registro dos procedimentos de busca – incluindo data, hora, quantidades, parâmetros, possíveis erros –, arquivos no formato CSV com todos os dados e os arquivos JPG das imagens baixadas (quando permitido pela Hemeroteca Digital Brasileira). É importante esmiuçar esses resultados.

O programa cria um diretório para armazenamento intitulado <HDB/{termo da busca}/{data da busca}/>. Nele é criado o diretório </CSV>, onde é salvo o arquivo CSV final contendo os seguintes dados para cada ocorrência: *Termo da busca, Data da Busca, Acervo, Ano, Edição, Página, Nome do arquivo, Link*. "CSV" significa "valores separados por vírgula" e é um formato muito comum para armazenar dados tabulares. Como nos explicam Folgert Karsdorp e colaboradores,

> É usado para armazenar informações tabulares de maneira semelhante a uma planilha. Em sua forma mais simples, cada linha em um arquivo CSV representa uma entrada de dados individual, sendo os atributos dessa entrada listados em uma série de campos separados por um delimitador (por exemplo, uma vírgula) (KARSDORP; KESTEMONT; RIDDELL, 2021, p. 36, tradução nossa).

Outro diretório criado é o <RELATÓRIOS>, onde são salvos os relatórios em arquivos de texto simples (em formato .txt) para cada busca e um arquivo em formato .csv com dados gerais dos acervos. Diferentes relatórios são produzidos e armazenados nesse diretório:

1. <GERAL_{termo da pesquisa}_{data e hora da busca}.txt> Contendo os dados gerais da pesquisa, inclui os seguintes campos: *Data e hora da busca; Local da busca; Período da busca; Periódico da busca; Termo da busca; Lista de acervos com ocorrências (máx. de 100); Quantidade de ocorrências; Total de acervos com ocorrências (máx. de 100); Total de páginas pesquisadas; Total de acervos pesquisados; Total de ocorrências; e Frequência de ocorrências por página*.

2. <relatório_{nome do acervo}_{número do acervo}_{data e hora da busca}.txt> Com os dados de cada acervo raspado, inclui os seguintes campos: *Data e hora da busca; Termo da busca; Acervo; Total de ocorrências; Link da lista de resultados*.

3. <ERRO_{número do acervo}_{data e hora da busca}.txt> Registra acervos que não puderam ser raspados. Contém estes dados: *Data e hora da busca; Termo da busca; Acervo com ocorrência que não pode ser acessado*.

Por fim, é criado um CSV com dados gerais relativo ao quantitativo de acervos e páginas pesquisadas na busca e à quantidade de ocorrências, assim como à frequência de ocorrências por página. Este arquivo conta com os seguintes dados: *Acervo, Total de Páginas, Total de Ocorrências, Frequência de ocorrências X páginas.*

No diretório <ARQUIVOS_IMG> são salvos os arquivos de imagens, no formato .jpg, das páginas com ocorrências, quando permitido pela Hemeroteca Digital Brasileira. O nome do arquivo é a combinação do número do acervo com o número geral da página (ex.: <168319_02_12603.jpg>).

Portanto, a estrutura de diretórios gerada pela ferramenta *pyHDB* pode ser representada da seguinte forma (Quadro 1):

Quadro 1 - Representação da estrutura de diretórios resultantes da *pyHDB*

```
HDB
├── {termo da busca}
│   ├── {data}
│   │   ├── CSV
│   │   ├── RELATÓRIOS
│   │   ├── ARQUIVOS_IMG
```

Fonte: Elaborado pelo autor a partir de Brasil (2021).

A partir do entendimento dessa estrutura técnica e metodológica, podemos refletir com mais detalhes sobre os resultados produzidos pela *pyHDB*.

Possibilidades analíticas

Neste último tópico, gostaria de apresentar três exemplos de buscas efetuadas por meio da *pyHDB*: uma busca com número de ocorrência na casa das dezenas, outra na casa das centenas e uma terceira na casa dos milhares. A primeira, realizada no dia 8 de outubro de 2021, reuniu os seguintes parâmetros:

- Local da busca: RJ;

- Período da busca: Todos;

- Periódico da busca: Todos;

- Termo da busca: "germano lopes da silva".

Esta combinação de parâmetros retornou 24 acervos com pelo menos uma ocorrência, totalizando 47 ocorrências. A *pyHDB* percorreu todas essas ocorrências, salvando os dados de cada uma delas para o arquivo CSV, criando os diretórios e fazendo download das imagens quando permitido. Todo esse processo foi completado num intervalo de 15 minutos e nove segundos, o que significa uma média de 19 segundos por ocorrência.

Germano Lopes da Silva, importante líder carnavalesco da cidade do Rio de Janeiro nas duas primeiras décadas do século XX, era genro de Tia Ciata e participou ativamente da fundação e manutenção de clubes carnavalescos, ao mesmo tempo em que era bedel da Escola Politécnica da cidade, membro da Guarda Nacional e eleitor registrado para votar (BRASIL, 2016, 2018). Sua presença nos jornais esteve quase sempre atrelada ao carnaval e suas festas, mas é possível encontrá-lo ao longo de todo o ano atuando na vida pública da cidade.

Apesar disso, sua presença nas páginas impressas dos jornais não passa da casa das dezenas. Isto prova que é possível lidar de forma adequada com as fontes sem o uso de recursos computacionais mais complexos. Ainda assim, é importante destacar que a repetição manual da coleta desses dados demanda tempo, muita atenção e favorece a incidência de erros no processo. Por conseguinte, o uso da *pyHDB* para esse caso é útil e recomendável. Primeiramente, é possível recuperar todas as ocorrências com facilidade, incluindo as informações básicas sobre cada uma delas (jornal, edição, página, ano, termo da busca, data da busca, nome do arquivo de imagem associado à ocorrência e o *hyperlink*) a partir do CSV gerado (Figura 2). Portanto, a instância de

análise, da leitura atenta de cada ocorrência, é operada a partir de um documento bem estruturado que permite acessar e trilhar todas essas ocorrências minimizando erros e lacunas.

Figura 2 - Print de visualização do arquivo CSV com os dados da busca do exemplo 2, mostrando as 16 primeiras linhas e todas as colunas

	Termo da busca	Data da Busca	Acervo	Ano	Edição	Página	Nome do arquivo	Link
0	"hemeterio jose dos santos"	2021-10-03	Gazeta de Noticias (RJ) - 1890 a 1899	1890	223	2	103730_03_1278.jpg	http://memoria.bn.br/docreader/103730_03/1278
1	"hemeterio jose dos santos"	2021-10-03	Gazeta de Noticias (RJ) - 1890 a 1899	1892	65	1	103730_03_5309.jpg	http://memoria.bn.br/docreader/103730_03/5309
2	"hemeterio jose dos santos"	2021-10-03	Gazeta de Noticias (RJ) - 1890 a 1899	1894	146	1	103730_03_9915.jpg	http://memoria.bn.br/docreader/103730_03/9915
3	"hemeterio jose dos santos"	2021-10-03	Diario de Noticias (RJ) - 1885 a 1895	1888	997	1	369365_4047.jpg	http://memoria.bn.br/docreader/369365/4047
4	"hemeterio jose dos santos"	2021-10-03	Diario de Noticias (RJ) - 1885 a 1895	1890	773	2	369365_7190.jpg	http://memoria.bn.br/docreader/369365/7190
5	"hemeterio jose dos santos"	2021-10-03	Diario de Noticias (RJ) - 1885 a 1895	1892	429	1	369365_10331.jpg	http://memoria.bn.br/docreader/369365/10331
6	"hemeterio jose dos santos"	2021-10-03	Jornal do Commercio (RJ) - 1940 a 1949	1948	304	6	nan	http://memoria.bn.br/docreader/364568_13/42282
7	"hemeterio jose dos santos"	2021-10-03	A Manhã (RJ) - 1925 a 1953	1927	319	2	116408_2062.jpg	http://memoria.bn.br/docreader/116408/2062
8	"hemeterio jose dos santos"	2021-10-03	A Manhã (RJ) - 1925 a 1953	1946	622	4	116408_32108.jpg	http://memoria.bn.br/docreader/116408/32108
9	"hemeterio jose dos santos"	2021-10-03	A Lanterna : Jornal da Noite (RJ) - 1916	1916	18	6	211702_110.jpg	http://memoria.bn.br/docreader/211702/110
10	"hemeterio jose dos santos"	2021-10-03	O Jornal (RJ) - 1940 a 1949	1949	16	5	nan	http://memoria.bn.br/docreader/110523_04/51090
11	"hemeterio jose dos santos"	2021-10-03	Almanak Henault (RJ) - 1909 a 1911	1909	1	344	709930_344.jpg	http://memoria.bn.br/docreader/709930/344
12	"hemeterio jose dos santos"	2021-10-03	Brazil Moderno (RJ) - 1906 a 1921	1908	16	32	101044_852.jpg	http://memoria.bn.br/docreader/101044/852
13	"hemeterio jose dos santos"	2021-10-03	Brazil Moderno (RJ) - 1906 a 1921	1908	16	41	101044_861.jpg	http://memoria.bn.br/docreader/101044/861
14	"hemeterio jose dos santos"	2021-10-03	Brazil Moderno (RJ) - 1906 a 1921	1908	16	44	101044_864.jpg	http://memoria.bn.br/docreader/101044/864
15	"hemeterio jose dos santos"	2021-10-03	Diario do Brazil (RJ) - 1881 a 1885	1883	170	3	225029_2365.jpg	http://memoria.bn.br/docreader/225029/2365

Fonte: Elaborado pelo autor a partir de Brasil (2021).

A recuperação de cada informação desse processo também é de grande importância para as pesquisas em história e as publicações resultantes delas, pois cada fonte utilizada pode ser referenciada de forma correta e precisa, inclusive com o link único para aquela fonte específica.

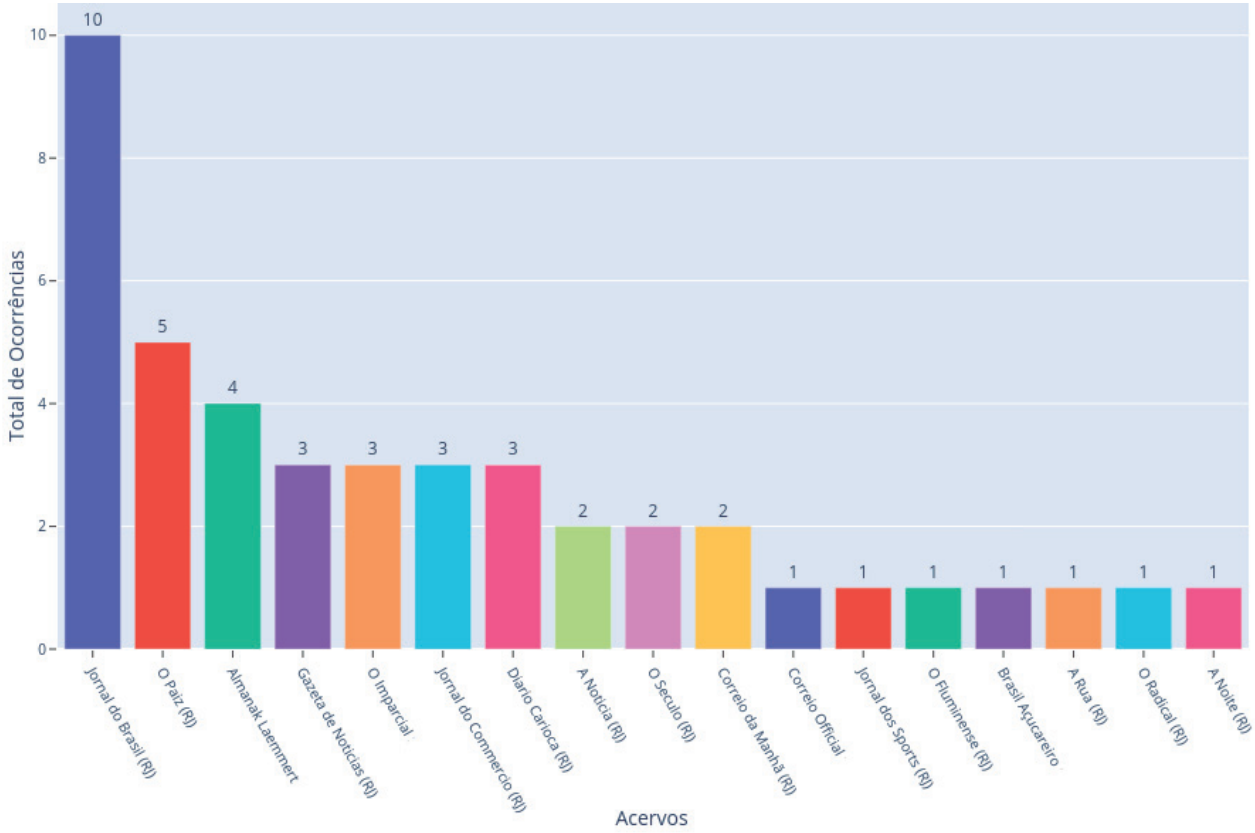
Outra possibilidade é a construção de visualizações dos dados referentes à busca de acordo com nossas necessidades de pesquisa e perguntas específicas. Por exemplo, no caso que acabamos de citar é interessante observar a concentração de ocorrências no *Jornal do Brasil* (Figura 3), importante veículo de divulgação do carnaval carioca na época.[4]

E também a grande concentração de ocorrências na década de 1900, momento da formação dos principais clubes carnavalescos onde Germano atuou – o Rosa Branca e o Macaco é Outro –, e na década de 1920, com a consolidação do samba e dos concursos carnavalescos como um dos principais símbolos do Rio de Janeiro (Figura 4).

Não há dúvidas de que essas visualizações e as novas perguntas que surgem delas e que elas possibilitam têm um papel restrito em buscas com apenas algumas dezenas de ocorrências. De fato, isso muda significativamente no caso de resultados na casa das centenas e milhares, como veremos nos próximos dois exemplos.

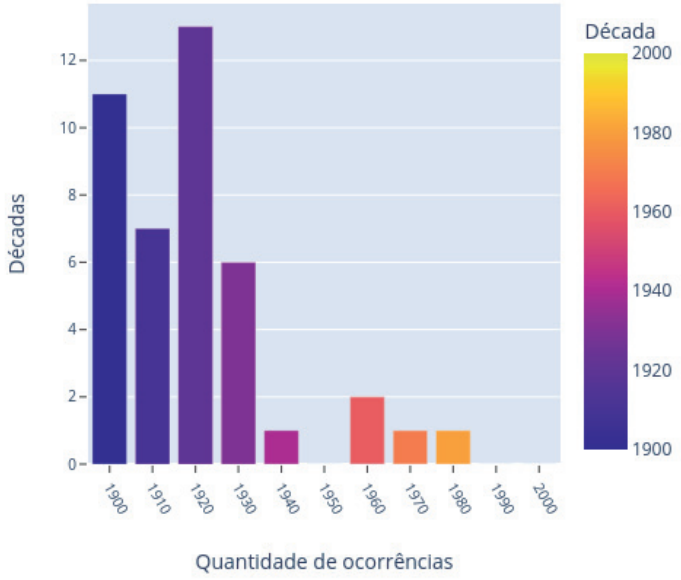
[4] Todos os gráficos estão disponíveis on-line com recursos interativos no link: https://ericbrasiln.github.io/analise_pyHDB/graficos .

Figura 3 - Gráfico de barras mostrando a quantidade de ocorrências absolutas por periódico



Fonte: Elaborado pelo autor a partir de Brasil (2021).

Figura 4 - Gráfico de barras mostrando a quantidade de ocorrências absolutas por década



Fonte: Elaborado pelo autor a partir de Brasil (2021).

A partir de determinado volume de ocorrências, a tarefa de coletar todos os dados manualmente se torna humanamente inviável. Vejamos o exemplo da busca realizada no dia 3 de outubro às 22h04min, cujos parâmetros foram *RJ* como local, todos os períodos e "hemeterio jose dos santos" como termo de busca.

O professor de língua portuguesa Hemetério José dos Santos atuou como educador na cidade do Rio de Janeiro nas primeiras décadas do século XX. Sua atuação política na cidade pautava a educação antirracista e promoveu importantes reflexões de enfrentamento ao racismo científico estruturante das relações sociais da sociedade brasileira. As pesquisas de Luara Santos tem possibilitado a compreensão de sua trajetória, assim como do papel de professores e professoras negras na cidade durante a Primeira República (SANTOS, 2015).

A busca por seu nome retornou 630 ocorrências, que precisaram de três horas, 33 minutos e 22 segundos para serem raspadas. Nesse ritmo, temos uma média de nove segundos por ocorrência.[5] Manualmente, se salvássemos todos os dados de cada ocorrência em um CSV e baixássemos a imagem em JPG, levaríamos uma média de 60 segundos para completar as ações. Ou seja, levaria mais de 10 horas de trabalho atingirmos o mesmo resultado!

Contudo, afirmo que esse tipo de ferramenta não deve ser encarado apenas como um "acelerador" da pesquisa. A *pyHDB* garante a coleta e organização dos dados de forma padronizada, atendendo aos interesses específicos do pesquisador; possibilita a visualização desses dados também de forma organizada; permite retomar cada ocorrência no momento necessário para sua leitura e análise; e estabelece o registro dos caminhos da pesquisa, garantindo o rigor e a transparência metodológica do trabalho. Cria um *subcorpus* documental a partir do oceano de dados disponibilizados pela Hemeroteca Digital Brasileira.

Esse ato de criar um *corpus* delimitado a partir de acervos digitais tem sido uma prática cada vez mais necessária para a pesquisa na era digital. Como demonstram Pfanzelter e colaboradores (2020), parte significativa das interfaces de hemerotecas digitais europeias também cria essa demanda para usuários que pretendem utilizar seus acervos para objetivos que transcendem a lista simples de resultados: "Sem a possibilidade de criarem *subcorpora*, os pesquisadores de humanidades muitas vezes não conseguem alinhar suas análises com suas questões de pesquisa específicas" (PFANZELTER *et al.*, 2020, p. 2, tradução nossa).

[5] É importante lembrar que esses dados relativos ao tempo variam muito em função do horário de acesso, da velocidade de resposta do servidor da Hemeroteca Digital Brasileira, dos possíveis erros gerados nessas respostas, da velocidade de banda da internet do usuário. Na nossa pesquisa, procuramos sempre executar o programa durante a noite para diminuir possíveis sobrecargas do servidor. A média de velocidade de banda utilizada foi de 60MB, em uma máquina com CPU AMD Ryzen 5 3400G (8) 3.700GHz, memória RAM de 8GB e sistema operacional Pop!_OS 21.04 x86_64.

Ehrmann e colaboradores, em pesquisa com 24 interfaces de hemerotecas digitais, apontam que há uma "lacuna entre as expectativas crescentes do usuário, estimuladas pelos avanços da mineração de texto, e as capacidades atuais de interface" (EHRMANN; BUNOUT; DÜRING, 2019, p. 17, tradução nossa).

A ferramenta *pyHDB* possibilita que pesquisadores implementem diversas técnicas de análise e visualização ligadas às humanidades digitais a partir de seus resultados: visualização em linhas do tempo, percepção do volume de ocorrências em períodos específicos; localização no espaço (dependendo do conjunto de jornais pesquisados).

A ferramenta, partindo dos dados sobre o número total de ocorrências de cada acervo e o número absoluto de páginas, calcula a proporção de quantas páginas são necessárias para o retorno de uma ocorrência. Isto nos permite avaliar o peso proporcional das ocorrências em cada acervo e não apenas seu número absoluto. Conforme a tabela a seguir (Figura 5), o professor Hemetério José dos Santos esteve representado com muito mais frequência no periódico *A Escola*, aparecendo seu nome impresso pelo menos uma vez a cada 49 páginas digitalizadas do acervo. Já no *Almanak Laemmert*, onde encontramos o maior número absoluto de ocorrências, o nome de Hemetério é encontrado apenas a cada 1.700 páginas.

Esse tipo de análise originada pelo cálculo de proporcionalidade de ocorrências em relação ao total de páginas possibilita a compreensão e a avaliação de relevância para determinados acervos que, por terem um número absoluto reduzido, poderiam passar despercebidos na análise.

Para buscas com milhares de ocorrências, as possibilidades são sobremaneira expandidas. Nosso terceiro exemplo retornou 5.335 ocorrências com os seguintes parâmetros: local, RJ; período, 1900-1909 e 1910-1919; termo da busca: "monteiro lopes". Neste caso, foram realizadas duas pesquisas, cada uma para uma década, em função do elevado número de ocorrências. Ambas as pesquisas foram realizadas no dia 06 de outubro de 2021.

Manoel da Motta Monteiro Lopes, advogado com doutorado pela Faculdade de Direito de Recife, chegou ao Rio de Janeiro no ano de 1894 e, em 1903, foi eleito membro do Conselho Municipal. Cumpriu seu mandato (1903-1904) e concorreu à reeleição. Mas, apesar da votação significativa, não foi reconhecido como eleito, fato que se repetiu nas eleições de 1905 para deputado federal. Em 1909, se candidatou novamente a deputado federal, mas dessa vez pelo Partido Republicano Democrata, e foi eleito. A ameaça de não reconhecimento de sua eleição na Câmara de Deputados mobilizou entidades formadas por trabalhadores negros de todo o Brasil. Em abril do mesmo ano, Monteiro Lopes foi diplomado deputado federal. Não completou o mandato, pois faleceu no final do ano de 1910 (DANTAS, 2010).

Figura 5 - Tabela com a variação de valores absolutos e proporcionais de ocorrências nos 10 primeiros acervos

Acervo	Ocorrências	Frequências	Variação
Almanak Laemmert : Administrativo, Mercantil e Industrial (RJ) - 1891 a 1940	1	20	-19
O Paiz (RJ) - 1910 a 1919	2	7	-5
O Tempo (RJ) - 1891 a 1894	3	2	1
Jornal do Brasil (RJ) - 1900 a 1909	4	17	-13
Gazeta de Noticias (RJ) - 1900 a 1919	5	21	-16
Annaes da Camara dos Deputados (RJ) - 1900 a 1910	6	25	-19
A Noticia (RJ) - 1894 a 1916	7	12	-5
A Escola : Revista Brasileira de Educação e Ensino (RJ) - 1877 a 1878	8	1	7
A Imprensa (RJ) - 1898 a 1914	9	13	-4
A Epoca (RJ) - 1912 a 1919	10	15	-5

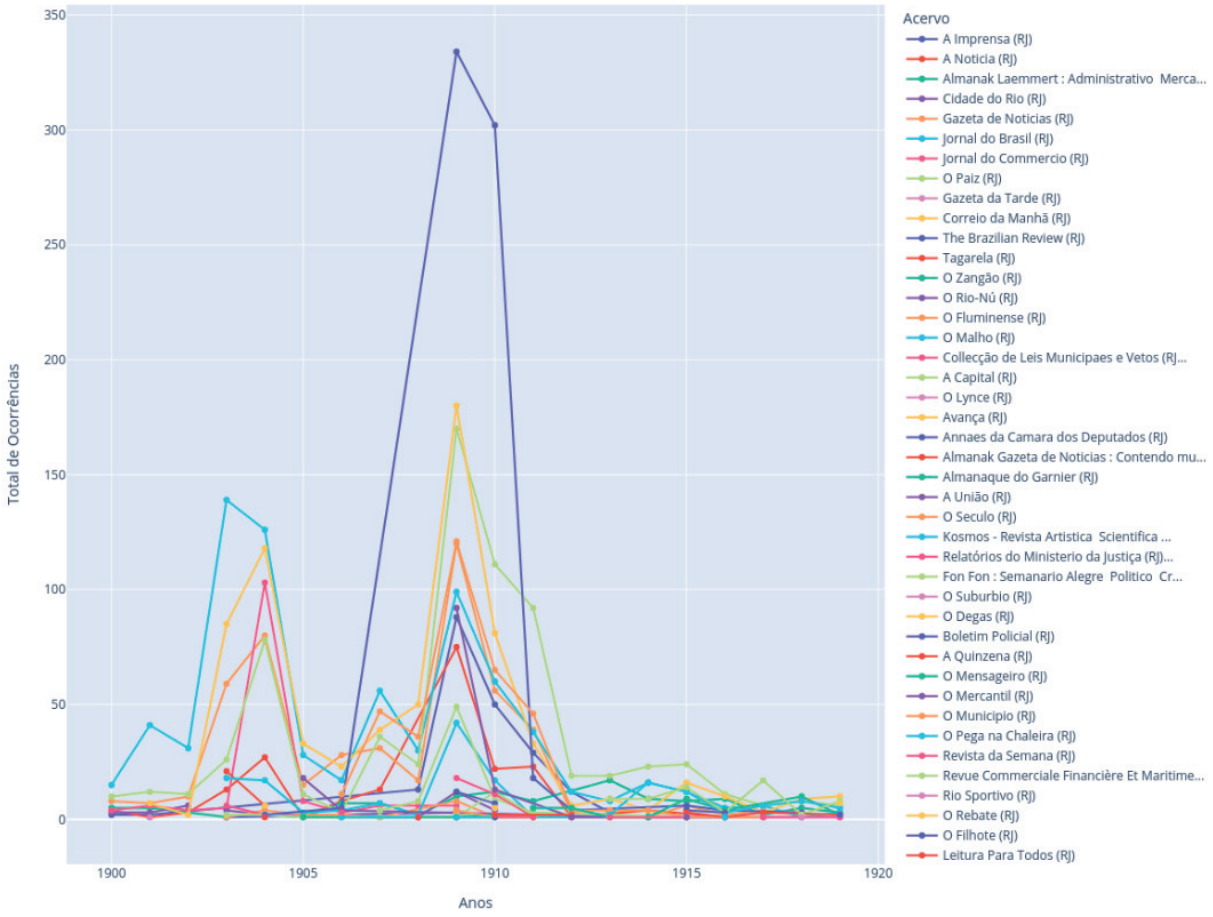
Fonte: Elaborado pelo autor a partir de Brasil (2021).

Os estudos de Carolina Vianna Dantas analisam como a atuação política de Monteiro Lopes e a mobilização popular em torno de sua candidatura apontam para as estratégias de ação política de trabalhadores negros na Primeira República e para os obstáculos criados pela estrutura racista dessa mesma república para o exercício da cidadania (DANTAS, 2010).

Analisando os resultados da busca por "monteiro lopes", podemos construir uma visualização da quantidade de ocorrências em cada acervo ao longo do tempo (Figura 6). Os dados apontam justamente para uma presença destacada do nome em dois períodos: o primeiro, entre 1903 e 1905, período de seu mandato no Conselho Municipal e das eleições em que sua

diplomação não foi reconhecida; o segundo se localiza entre 1909 e 1910, quando ocorreu o intenso debate acerca da candidatura e de seu reconhecimento ou não como deputado federal (no ano de 1909) e sua curta atuação como deputado até final de 1910. Em relação a este último momento, é importante destacar o alto número de ocorrências nos *Annaes da Câmara dos Deputados (RJ),* o que aponta para sua intensa participação política, mesmo que por um curto período.

Figura 6 - Gráfico de linha com as ocorrências organizadas por anos e jornais



Fonte: Elaborado pelo autor a partir de Brasil (2021).

O objetivo deste artigo não é realizar a análise qualitativa desses resultados nem seus gráficos – os exemplos usados, inclusive, foram propositalmente escolhidos a partir de pesquisas já publicadas e buscaram corroborar as hipóteses de seus autores (BRASIL, 2016; DANTAS, 2010; SANTOS, 2015) – mas apresentar possibilidades do uso da *pyHDB* para diferentes conjuntos de dados.

O *data set* utilizado neste artigo, gerado pela ferramenta *pyHDB* a partir das buscas dos termos utilizados como exemplo, está disponível no repositório do Zenodo e identificado pelo DOI 10.5281/zenodo.6003276. Os gráficos foram gerados com a biblioteca *Plotly* a partir das análises dos CSV utilizando-se *scripts* escritos em Python e a biblioteca *Pandas*, especializada em análise de dados (MCKINNEY, 2018). Esses *scripts* (trechos de códigos) e os gráficos em formato interativo – salvos em arquivo .html – estão disponíveis no repositório do GitHub ericbrasiln/analise_pyHDB e na página web https://github.com/ericbrasiln/pyHDB.

Sem dúvida, com mais pesquisadores e pesquisadoras utilizando a ferramenta *pyHDB*, endereçando seus próprios interesses e perguntas por meio dela, novas possibilidades e características poderão ser incluídas no código-fonte. Sugestões, correções e ampliações são muito bem-vindas e encorajadas.[6] Este é só o começo.

Considerações finais

A ferramenta *pyHDB* é ao mesmo tempo um exemplo e um esforço para construir procedimentos metodológicos rigorosamente registrados e controlados para todos e todas os que usam a interface da Hemeroteca Digital Brasileira. É também um passo em direção à transparência e ao rompimento com a opacidade desses procedimentos de pesquisa aplicados pelos usuários nas interfaces de busca de acervos digitais.

A ferramenta proporciona à comunidade de pesquisadores e usuários da Hemeroteca Digital Brasileira em geral um poderoso companheiro de pesquisa, mas também de análise dos resultados, enquanto a própria interface não disponibiliza recursos mais sofisticados para os usuários. Para que isto ocorra, ao meu ver, será necessária uma política consistente e de longo prazo de investimentos por parte do poder público para a realização de projetos em parceria com as instituições arquivísticas e universidades, centros e laboratórios de diferentes disciplinas, para que novas ferramentas sejam disponibilizadas pela própria interface: opções de acesso ao arquivo txt com o texto do OCR; opções de XML, acesso ao arquivo Dublin Core; dados relativos à frequência de palavras; precisão do OCR em cada página; desenvolvimento de um espaço de trabalho para os usuários; geração de relatórios de busca; entre outras. Isto não deve isentar a instituição e toda a sociedade do debate sobre a transparência metodológica em todas as etapas de projetos financiados com dinheiro público.

[6] É possível criar *issues* e *pull requests* diretamente no repositório do programa no GitHub: https://github.com/ericbrasiln/pyHDB.

Ao escrever o código de uma ferramenta dessa natureza, precisei conhecer a fundo os limites, possibilidades e inter-relações entre códigos, acervo e resultados de pesquisa. Pude compreender tais estruturas e escrever um código que atendesse a demandas específicas de pesquisa. Compreendo que o desenvolvimento de uma ferramenta com essas características exige conhecimentos muito especializados e que sua curva de aprendizado é bem mais elevada do que a envolvida em programas com interface gráfica para usuários. Apesar de ter buscado documentar todas as etapas de funcionamento da *pyHDB*, registrando-as na documentação disponível no repositório público e imprimindo na tela detalhes e o progresso de uso, reconheço que a ausência de uma interface gráfica é uma barreira. Alternativas como a criação de uma interface de linha de comando (CLI) ou um conjunto de *Jupyter Notebooks* poderiam possibilitar a ampliação de seu uso, não estando descartadas para o futuro. Entretanto, isso demandaria um investimento de tempo e uma dedicação muito maiores, sobretudo em um projeto sem financiamento como este.

As possibilidades de coleta e organização dos dados dependem da construção do problema e de interesses de pesquisa do investigador em constante diálogo com aspectos tecnológicos e arquivísticos. Como afirma Marres e Weltevrede, a raspagem " parece implicar uma abordagem distinta à produção de conhecimento. Indiscutivelmente vem com uma epistemologia embutida" (MARRES; WELTEVREDE, 2013, p. 319, tradução nossa). Por isso é e será cada vez mais relevante a compreensão mínima do funcionamento dos algoritmos que estruturam as ferramentas utilizadas na pesquisa.

Especificamente no caso da história, concordo com Ian Milligan quanto a que "nem todos os/as historiadores/as tornar-se-ão programadores. Ao invés disso, eles precisam ser capazes de implementar – com compreensão – algoritmos criados por outras pessoas" (MILLIGAN, 2019, p. 155, tradução nossa). O distanciamento entre historiadores/as e a elaboração das ferramentas de pesquisa e os conjuntos de dados disponibilizados atualmente precisa ser um alerta para o risco de perdemos a capacidade de criticar essas fontes e métodos.

Pensar algoritmicamente — ou seja, compreender as estruturas e lógicas que pautam e definem os passos encadeados que determinado programa ou ferramenta vai trilhar – permite o desenvolvimento de práticas de heurística digital e de reflexões teóricas acerca dos próprios resultados atingidos na interação entre pesquisador, ferramenta e fontes digitais. Escrever as centenas de linhas do código que animam a ferramenta *pyHDB* colocaram esse desafio na tela do meu computador de forma escancarada, reforçando sua relevância para a história na era digital.

Referências

BENJAMIN, Ruha. Assessing risk, automating racism. **Science**, New York, v. 366, n. 6464, p. 421–422, 25 out. 2019a. Disponível em: https://doi.org/10.1126/science.aaz3873. Acesso em 1 out. 2021.

BENJAMIN, Ruha. **Race after technology**: abolitionist tools for the new Jim code. Cambridge: Wiley, Polity Press, 2019b.

BETTENCOURT, Angela Maria Monteiro; PINTO, Monica Rizzo Soares. A hemeroteca digital brasileira. In: CONGRESSO BRASILEIRO DE BIBLIOTECONOMIA, DOCUMENTAÇÃO E CIÊNCIA DA INFORMAÇÃO, XXV, 2013, Florianópolis. **Anais** [...], Florianópolis: FEBAB, 2013, p. 1028–1038.

BIRHANE, Abeba. Algorithmic injustice: a relational ethics approach. **Patterns**, Amsterdam, v. 2, n. 2, p. 1-9, 12 fev. 2021. https://doi.org/10.1016/j.patter.2021.100205. Acesso em: 12 set. 2022.

BRASIL, Eric. **Carnavais Atlânticos:** Cidadania e Cultura Negra no pós-abolição do Rio de Janeiro, Brasil e Porto de Espanha, Trinidad (1838-1920). 2016. Tese (Doutorado em História), Universidade Federal Fluminense, Niterói, 2016.

BRASIL, Eric. **pyHDB: ferramenta heurística para a Hemeroteca Digital Brasileira**. Zenodo, 2021. Disponível em: https://zenodo.org/record/5706507. Acesso em: 12 set. 2022.

BRASIL, Eric. Germano Lopes da Silva: experiências de um carnavalesco, eleitor e cidadão no Distrito Federal (c. 1900-1930). 2018. **Biblioteca Consuelo Pondé**. Disponível em: http://www.bvconsueloponde.ba.gov.br/modules/conteudo/conteudo.php?conteudo=201. Acesso em: 12 set. 2022.

BRASIL, Eric; NASCIMENTO, Leonardo Fernandes. História digital: reflexões a partir da Hemeroteca Digital Brasileira e do uso de CAQDAS na reelaboração da pesquisa histórica. **Revista Estudos Históricos**, Rio de Janeiro, v. 33, n. 69, p. 196–219, 1 jan. 2020. Disponível em: http://dx.doi.org/10.1590/S2178-14942020000100011. Acesso em: 12 set. 2022.

BRESCIANO, Juan Andrés. **La investigación histórica y las nuevas tecnologías**. Montevideo: Librería de la Facultad de Humanidades y Ciencias de la Educación, 2000.

CLAVERT, F.; FICKERS, A. On pyramids, prisms, and scalable reading. **Journal of Digital History**, jdh001, 2021. Disponível em: https://www.journalofdigitalhistory.org/en/article/jXupS3QAeNgb. Acesso em: 12 set. 2022.

COSTA, Marcela Albaine. **Ensino de história e historiografia escolar digital**. 1. ed. Curitiba: EDITORA CRV, 2021. DOI 10.24824/978655868256.1.

DANTAS, Carolina Vianna. Monteiro Lopes (1867-1910), um "líder da raça negra" na capital da república. **Afro-Ásia**, Salvador, n. 41, 2010. p. 167-209. DOI 10.9771/aa.v0i41.21201. Disponível em: https://periodicos.ufba.br/index.php/afroasia/article/view/21201. Acesso em: 12 set. 2022.

EHRMANN, Maud; BUNOUT, Estelle; DÜRING, Marten. Historical Newspaper User Interfaces: A Review. In: LIBRARIES: DIALOGUE FOR CHANGE. Atenas, 2017. **Anais** [...], Atenas: IFLA WLIC, 2019, p. 1-24. Disponível em: http://library.ifla.org/id/eprint/2578/. Acesso em: 14 set. 2021.

FRIDLUND, Mats; OIVA, Mila; PAJU, Petri (org.). **Digital Histories**: Emergent Approaches within the New Digital History. Helsinki: Helsinki University Press, 2020.

FUNDAÇÃO BIBLIOTECA NACIONAL. **Por motivos técnicos, nosso site se encontra temporariamente fora do ar. Esperamos solucionar o problema o quanto antes**. Rio de Janeiro, 12 abr. 2021. Twitter: @FBN. Disponível em: http://pic.twitter.com/OUGMDWE3hJ. Acesso em: 22 out. 2021.

GALLOWAY, Alexander R. **The Interface Effect**. Cambridge: Polity, 2012.

GOODING, Paul; TERRAS, Melissa; WARWICK, Claire. The myth of the new: mass digitization, distant reading, and the future of the book. **Literary and Linguistic Computing**, Oxford, v. 28, n. 4, p. 629–639, dez. 2013. DOI: http://dx.doi.org/10.1093/llc/fqt051. Acesso em: 12 set. 2022.

Impresso. **Media Monitoring of the Past**. Supported by the Swiss National Science Foundation under grant CR- SII5_173719, 2019. Disponível em: https://impresso-project.ch. Acesso em: 12 set. 2022.

JARLBRINK, Johan; SNICKARS, Pelle. Cultural heritage as digital noise: nineteenth century newspapers in the digital archive. **Journal of Documentation**, Bingley, v. 73, n. 6, p. 1228–1243, 12 out. 2017. Disponível em: https://doi.org/10.1108/JD-09-2016-0106. Acesso em: 12 set. 2022.

JENSEN, Helle Strandgaard. Digital Archival Literacy for (all) Historians. **Media History**, Londres, v. 27, n. 2, p. 251–265, 2021. Disponível em: https://doi.org/10.1080/13688804.2020.1779047. Acesso em: 12 set. 2022.

KARSDORP, Folgert; KESTEMONT, Mike; RIDDELL, Allen. **Humanities data analysis**: case studies with Python. Princeton: Princeton University Press, 2021.

KEMMAN, Max. **Trading Zones of Digital History**. Berlin: De Gruyter Oldenbourg, 2021. DOI 10.1515/9783110682106.

KROTOV, Vlad; JOHNSON, Leigh; SILVA, Leiser. Tutorial: Legality and Ethics of Web Scraping. **Communications of the Association for Information Systems**, Atlanta, v. 47, n. 1, p. 539–563, 2020. Disponível em: https://doi.org/10.17705/1CAIS.04724. Acesso em: 12 set. 2022.

LUCCHESI, Anita. **Digital history e Storiografia digitale**: estudo comparado sobre a escrita da história no tempo presente (2001-2011). 2014. Dissertação (Mestrado em História), Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2014.

MARRES, Noortje; WELTEVREDE, Esther. Scraping the Social? **Journal of Cultural Economy**, Londres. v. 6, n. 3, p. 313–335, 2013. Disponível em: https://doi.org/10.1080/17530350.2013.772070. Acesso em: 12 set. 2022.

MCKINNEY, Wes. **Python Para Análise de Dados**: Tratamento de Dados com Pandas, NumPy e IPython. 1ª edição. São Paulo: Novatec Editora, 2018.

MILLIGAN, Ian. **History in the Age of Abundance?**: How the Web Is Transforming Historical Research. 328. ed. London; Chicago: McGill-Queen's University Press, 2019.

MITCHELL, Ryan. **Web Scraping with Python**: Collecting More Data from the Modern Web. 2. ed. Sebastopol, CA: O'Reilly Media, 2018.

NASCIMENTO, Leonardo F. **Sociologia digital**: uma breve introdução. Salvador: EDUFBA, 2020.

NASCIMENTO, Leonardo Fernandes. Combinando webscraping em R e ATLAS.ti na pesquisa em ciências sociais: as possibilidades e desafios da sociologia digital. In: CONGRESSO BRASILEIRO DE SOCIOLOGIA, 18, 2017, Brasília. **Anais** [...], Brasília: Sociedade Brasileira de Sociologia, 2017, p. 2-17.

NICHOLSON, Bob. The Digital Turn. **Media History**,Londres, v. 19, n. 1, p. 59–73, 1 fev. 2013. Disponível em: https://doi.org/10.1080/13688804.2012.752963. Acesso em: 12 set. 2022.

NICODEMO, Thiago Lima; CARDOSO, Oldimar Pontes. Meta-história para robôs (bots): o conhecimento histórico na era da inteligência artificial. **História da Historiografia**: International Journal of Theory and History of Historiography, Ouro Preto, v. 12, n. 29, 28 abr. 2019. DOI 10.15848/hh.v12i29.1443. Acesso em: 12 set. 2022.

NOBLE, Safiya Umoja. **Algorithms of oppression**: data discrimination in the age of Google. New York: New York University Press, 2018.

PFANZELTER, Eva; OBERBICHLER, Sarah; MARJANEN, Jani; et al. Digital interfaces of historical newspapers: opportunities, restrictions and recommendations. **Journal of Data Mining and Digital Humanities**, v. HistoInformatics, jdmdh:6121, 2021. Disponível em: https://jdmdh.episciences.org/7069. Acesso em: 12 set. 2022.

PIROVANI, Juliana; OLIVEIRA, Elias. Portuguese named entity recognition using conditional random fields and local grammars. 2018. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC 2018), 11, 2018, Miyazaki. **Anais** [...]. Miyazaki: European Language Resources Association (ELRA), 2018, p. 4452-4456.

Programming Historian, ISSN: 2397-2068. Disponível em: https://programminghistorian.org. Acesso em: 12 set. 2022.

RAYMOND, Eric S. **The art of Unix programming**. Harlow: Addison-Wesley, 2003.

RÖHLE, Bernhard Rieder Theo; RIEDER, Bernhard. Digital Methods: Five Challenges. *In*: BERRY, David M. (org.). **Understanding Digital Humanities**. London: Palgrave Macmillan UK, 2012. p. 67–84. DOI 10.1057/9780230371934_4.

ROMEIN, C. Annemieke; KEMMAN, Max; BIRKHOLZ, Julie M.; BAKER, James; GRUIJTER, Michel De; MERO□O□PEÑUELA, Albert; RIES, Thorsten; ROS, Ruben; SCAGLIOLA, Stefania. State of the Field: Digital History. **History**, Hoboken, v. 105, n. 365, p. 291–312, 2020. Disponível em: https://doi.org/10.1111/1468-229X.12969. Acesso em: 12 set. 2022.

SALGANIK, Matthew J. **Bit by Bit**: Social Research in the Digital Age. Reprint edição. Princeton: Princeton University Press, 2017.

SALMI, Hannu. **What is Digital History?** 1ª edição. Cambridge: Polity, 2020. What is History? Series.

SANTOS, Luara. **'Etymologias preto':** Hemetério José dos Santos e as questões raciais de seu tempo (1888-1920). 2015. Dissertação (Mestrado em História), Centros Federais de Educação Tecnológica, Rio de Janeiro, 2015.

SHERRATT, Tim. **GLAM Workbench (version v1.0.0)**. Zenodo, 2021. Disponível em: https://doi.org/10.5281/zenodo.5603060. Acesso em: 12 set. 2022.

SILVA, Tarcízio. Racismo Algorítmico em plataformas digitais: microagressões e discriminação em código. *In:* SILVA, Tarcízio (org.). **Comunidades, Algoritmos e Ativismos Digitais**: Olhares Afrodiaspóricos. São Paulo: LiteraRUA, 2020.

SILVEIRA, Pedro Telles da. **História, técnica e novas mídias**: reflexões sobre a história na era digital. Tese (Doutorado em História), UFRGS, Porto Alegre, 2018. Disponível em: <https://lume.ufrgs.br/handle/10183/189249>. Acesso em: 27 set. 2021.

SINCLAIR, Stéfan; ROCKWELL. Geoffrey. **Voyant Tools**. Web. Disponível em: http://voyant-tools.org/. Acesso em: 12 set. 2022.

SOLBERG, Janine. Googling the Archive: Digital Tools and the Practice of History. **Advances in the History of Rhetoric**, Londres, v. 15, n. 1, p. 53–76, 1 jan. 2012. Disponível em: https://doi.org/10.1080/15362426.2012.657052. Acesso em: 12 set. 2022.

SOUTHERTON, Clare. Datafication. *In*: SCHINTLER, Laurie A.; MCNEELY, Connie L. (org.). **Encyclopedia of Big Data**. Cham: Springer International Publishing, 2020. p. 1–4. DOI 10.1007/978-3-319-32001-4_332-1. Acesso em: 12 set. 2022.

TURKEL, William J.; CRYMBLE, Adam. Understanding Web Pages and HTML. **Programming Historian**, Londres, 17 jul. 2012. Disponível em: https://programminghistorian.org/en/lessons/viewing-html-files. Acesso em: 8 jun. 2021.

WALSH, Melanie. **Introduction to Cultural Analytics & Python**, Version 1, 2021, Disponível em: https://doi.org/10.5281/zenodo.4411250. Acesso em: 12 set. 2022.

WEIBEL, Stuart. **Metadata: the Foundations of Resource Description**. d-lib magazine. Disponível em: https://www.dlib.org/dlib/July95/07weibel.html. Acesso em: 3 out. 2021.

Informações adicionais

Biografia profissional

Eric Brasil é doutor em História pela Universidade Federal Fluminense. Professor de História no Instituto de Humanidades e Letras da Universidade da Integração Internacional da Lusofonia Afro-brasileira (IHL-UNILAB). Autor do livro *A Corte em Festa: experiências negras em carnavais do Rio de Janeiro (1879-1888)*. Editor do periódico *Programming Historian* em português. Pesquisador do Laboratório de Humanidades Digitais da Universidade Federal da Bahia (LABHDUFBA). Membro do GT Nacional Emancipações e Pós-Abolição da Associação Nacional de História (ANPUH).

Endereço para correspondência

Campus dos Malês, Av. Juvenal Eugênio Queiroz, s/n, sala da coordenação, Baixa Fria, CEP.: 43900-000, São Francisco do Conde, BA, Brasil.

Financiamento

Não se aplica.

Agradecimentos

Agradeço aos membros do Laboratório de Humanidades Digitais da UFBA pelo espaço cedido para desenvolver pesquisas, especialmente a Gabriel Andrade, pelas dicas e ajuda com Python, e ao seu coordenador, Leonardo F. Nascimento, pela parceria, leitura e por ter sugerido o nome *pyHDB*. Agradeço a Yaci Farias e Luara Santos pela leitura e comentários durante a escrita do artigo. Agradeço a Ana Carolina Veloso e Priscila Valverde Silveira pelo trabalho desenvolvido como bolsistas de iniciação científica ao longo de 2020 e 2021. Agradeço a Alexandra Elbakyan por sua atuação na ciência livre.

Conflito de interesse

Não foi declarado nenhum conflito de interesse.

Aprovação no comitê de ética

Não se aplica.

Publicação prévia

Este artigo deriva da apresentação "Por uma Heurística Digital no ofício do Historiador", apresentado no evento II Congresso Internacional em Humanidades Digitais (2021).

Contexto da pesquisa

As reflexões contidas neste artigo estão inseridas no projeto de pesquisa "História Digital: acervos e ferramentas digitais para pesquisa e ensino", desenvolvido no IHL/UNILAB e nos debates propiciados pelo LABHDUFBA ao longo dos anos de 2020 e 2021.

Método de avaliação

Duplo-cega por pares.

Preprint

O artigo não é um preprint.

Disponibilidade de dados de pesquisa e outros materiais

Os conteúdos subjacentes ao artigo já estão disponíveis nos repositórios GitHub e Zenodo, disponíveis em https://github.com/ericbrasiln/pyHDB, https://zenodo.org/record/7032329, https://ericbrasiln.github.io/analise_pyHDB/, https://zenodo.org/record/6003276 .

Editores responsáveis

Flávia Varella - Editora-chefe

Fábio Joly - Editor responsável

Histórico de avaliação

Data de submissão: 14 de novembro de 2021

Data de alteração: 31 de janeiro de 2022

Data de aprovação: 18 de fevereiro de 2022

Direitos autorais

Licença