Softwares em humanidades digitais

Potencialidades e limitações metodológicas na pesquisa social com resgate de notícias da web

Softwares in digital humanities: methodological potentialities and limitations in social research with web news retrieval / Softwares en humanidades digitales: potencialidades y limitaciones metodológicas en investigación social con recuperación de noticias desde la web

Cristiane Naiara Araújo de Souza
Mestra em Ciências da Comunicação pela Universidade Federal do Amazonas (Ufam) e doutoranda em Comunicação e Informação pela Universidade Federal do Rio Grande do Sul (UFRGS), Brasil.
comunica.manaus.am@gmail.com

Karla Maria Muller
Doutora em Ciências da Comunicação pela Universidade do Vale do Rio dos Sinos (Unisinos). Professora do Departamento e do Programa de Pós-Graduação em Comunicação da UFRGS, Brasil.
kmmuller@ufrgs.br

RESUMO

Este artigo aborda potencialidades e limitações de softwares empregados na pesquisa social para coleta, sistematização, análise e visualização de dados com resgate de páginas na web (notícias). A abordagem qualitativa – pesquisa documental e análise de conteúdo – avalia 23 aplicações gratuitas retiradas do Laboratório em Rede de Humanidades Digitais, do Digital Humanities Lab e do SourceForge. Constatou-se a variedade de softwares capazes de oferecer soluções no tratamento medotológico com grande volume de dados heterogêneos, apesar das limitações quanto ao idioma e à literacia digital.

Palavras-chave: *softwares; pesquisa social; notícias da web; humanidades digitais.*

ABSTRACT

This article addresses the potentials and limitations of softwares used in social research to collect, systematize, analyze and visualize data with the retrieval of web pages (news). The qualitative approach – documental research and content analysis – evaluates 23 free applications extracted from the Laboratório em Rede de Humanidades Digitais, Digital Humanities Lab and SourceForge. It was found that a variety of softwares were capable of offering solutions in the methodological treatment of large volumes of heterogeneous data, although the limitations regarding language and digital literacy.

Keywords: *softwares; social research; web news; digital humanities.*

RESUMEN

Este artículo aborda los potenciales y las limitaciones de los softwares utilizados en la investigación social para recopilar, sistematizar, analizar y visualizar datos con páginas web (noticias). El enfoque cualitativo – investigación documental y análisis de contenido – evalúa 23 aplicaciones gratuitas del Laboratório em Rede de Humanidades Digitais, el Digital Humanities Lab y el SourceForge. Se encontró una variedad de softwares capaces de ofrecer soluciones en el tratamiento metodológico con grandes volúmenes de datos heterogéneos, a pesar de las limitaciones en cuanto al idioma y la alfabetización digital.

Palabras clave: *softwares; investigación social; noticias de la web; humanidades digitales.*

Introdução

A disponibilidade de fontes codificadas por meio da inteligência artificial e armazenadas no *big data* (Lewis; Westlund, 2015)[1] preconiza a atuação criteriosa dos pesquisadores para extrair dados relevantes, analisá-los e representar os resultados das investigações da forma adequada ao tipo de informação e ao escopo da pesquisa. Há de se considerar, sobretudo, que a existência de algoritmos, além da produção e da distribuição de conteúdos na web, torna a *artificial intelligence* e o *big data* dois dos fatores (atores) imprescindíveis em pesquisas com grande volume de dados.

Nesse sentido, antes mesmo de eleger um ou mais dos *softwares* disponíveis e capazes de contribuir para elucidar um objeto de investigação previamente eleito, cumpre dar um passo atrás para compreender as ferramentas ou as lentes (Liu, 2009) das quais os investigadores dispõem no contexto das humanidades digitais e na ambiência explorada. Outrossim, busca-se conhecer aplicações usadas na abordagem quali-quantitativa segundo as quais é possível, em alinhamento a pesquisas que tenham como objeto empírico um *corpus* de material jornalístico publicado na web, elucidar de modo coerente conteúdos obtidos nas etapas de coleta, sistematização, análise e visualização de dados com resgate de publicações em portais, mais particularmente um conjunto de cibernotícias sobre o qual o pesquisador se debruçará. De modo sistemático e organizado, será possível gerar compreensões sobre o tema, contribuindo para fortalecer a área da comunicação.

Importa mencionar que o objetivo de resgatar páginas noticiosas na web, além de realizar análises quantitativas e qualitativas combinadas, é a hipótese referencial neste texto tão somente em virtude da vinculação acadêmica das autoras à área do jornalismo. Ou seja, o objetivo aqui é compreender de que modo os *softwares* a seguir elencados podem ser instrumentalizados para o alcance desse virtual escopo de investigação, considerando um relativamente extenso e variado volume de dados na web. Sem embargo, é plenamente possível a combinação de aplicações para finalidades similares (não idênticas), tais como a pesquisa em repositórios de dados heterogêneos e multimidiáticos, como arquivos digitais de documentos públicos ou historiográficos etc.

Assim, antes de tentar encaixar o objeto numa ou em mais aplicações – sejam as criadas por centros de pesquisas de universidades e distribuídas

1 O que Lewis e Westlund (2015) propõem é um ponto de partida conceitual que aborda jornalismo em relação ao *big data*, onde é possível observar reações diversas dos produtores de conteúdos, como resistência e proatividade.

gratuitamente ou as elaboradas por empresas de tecnologia e com algum custo ou versões para estudantes, é importante conhecer um panorama mínimo de quais seriam eficazmente usadas em pesquisas cujo recorte se aproxima da proposta de viabilizar tratamento metodológico a notícias da web. Para tanto, crê-se ser relevante fazer o levantamento dos *softwares* mais efetivos quando empregados nas fases de coleta, sistematização, análise e visualização de dados, avaliando suas respectivas potencialidades e limitações.

Humanidades digitais, jornalismo e resgate de notícias na web

A professora Zeng (2017) apresenta o *smart data*, que, grosso modo, é o aprofundamento da disponibilidade do *big data*. Segundo afirma, "ele [*smart data*] fornece valor ao aproveitar desafios apresentados por volume, velocidade, variedade e veracidade do *big data*, fornecendo informações acionáveis e melhorando a tomada de decisão [...]" (p. 2). Os dados de diversas fontes são reunidos, relacionados e analisados para, então, "alimentar processos decisórios". Mas como transformar o *big data* em *smart data*? A seguir, são listadas estratégias para se alcançar um grande volume de dados confiáveis, contextualizados, relevantes, cognitivos, preditivos e consumíveis em qualquer escala (p. 3). Em suma, isso inclui:

> computação cognitiva, aprendizado profundo, aprendizado de máquina, inteligência artificial, análise preditiva, bancos de dados gráficos, inteligência de máquina, processamento de voz, tecnologias semânticas, veículos autônomos, *big data*, ciências de dados, internet das coisas (IoT), análise de texto, Resource Description Framework (RDF), gráficos de conhecimento, computação contextual, dados vinculados, raciocínio profundo, ontologias, JSON-LD (um Lightweight Linked Data Format), senso comum e processamento de linguagem natural (PNL) e pesquisa semântica. (Zeng, 2017, p. 3)

Tão importante quanto saber quais as ferramentas compõem o *smart data* é saber quem faz uso delas e com quais objetivos. Segundo Zeng (2017, p. 4), pesquisadores de diversas áreas já têm se arriscado na tarefa (engenheiros, cientistas naturais, analistas financeiros e até agentes públicos). Quanto ao jornalismo, especificamente em relação ao tipo de pesquisa ora ilustrado, o *corpus* se encaixaria na definição acima por se tratar de um conjunto de dados semiestruturados de acesso aberto e gratuito de formato similar (notícias na web disponíveis para não assinantes).

Segundo Aquino (2020), a ideia desenvolvida por Daniel Alves é a de que as humanidades digitais (HD) são uma comunidade de práticas. Portanto, ao invés de ser um campo institucionalizado de estudos acadêmicos (interdisciplinares), as HD se desenvolveriam mais fortemente a partir dos anos 2000 como uma comunidade de práticas que congrega investigadores cujos interesses são comumente materializados em parcerias e projetos (p. 3-4). Esse autor ainda prossegue com relevante observação crítica sobre as inúmeras aplicações dos softwares disponíveis no mercado:

> O conhecimento sobre programação ou como operar um Sistema de Informações Geográficas pode ser necessário a uma aplicação muito específica – por exemplo, eu tenho necessidade de dominar os SIGs; por outro lado, ainda não senti a necessidade de dominar um software de análise de redes, por não utilizar em meu trabalho. Mas eu sei o que o software faz, sei o que é possível obter daquela ferramenta. Isso me permite dialogar com quem trabalha com esse método, por exemplo, e não ficar "às escuras" à espera do que um algoritmo vai me mostrar. (Aquino, 2020, p. 13)

Em outras palavras, importa, mais do que saber operar todas essas ferramentas, extrair delas o melhor com que podem contribuir para o trabalho de pesquisa; e instrumentalizá-las em relação aos objetivos propostos, e não o contrário. Todavia, há vasta produção que necessita ser ao menos conhecida pelos pesquisadores das áreas para as quais são direcionadas. E mais ainda: podem sim ser elaboradas a partir da colaboração interdisciplinar de investigadores dos diferentes campos.

Quanto à colaboração entre pesquisadores e instituições, esta tem se mostrado eficaz para a criação de projetos bem sucedidos: Academia.edu e ResearchGate na categoria das redes sociais; GitHub (projetos open source), OSF (comunidades em várias áreas) e ArXiv (exatas e biológicas) como repositórios; e SciHub como alternativa às editoras e ao pagamento para aceder a trabalhos publicados em periódicos de acesso restrito. Nesses ambientes, além dos *papers* submetidos à revisão dos pares, aquela que antecede os aceites em periódicos acadêmicos e que está submetida a rígidas regras editoriais, são compartilhados trabalhos do tipo *preprint*. Cada tipologia tem suas vantagens e desvantagens, tema que, embora relevante, não está entre os objetivos deste artigo.

No Brasil, os repositórios universitários são ferramentas fundamentais para a abertura do conhecimento em todas as áreas, e ali "podem ser disponibilizadas dissertações, teses, *working papers*, artigos, livros, atas, *preprints*, de tudo, toda a nossa produção científica, inclusive conferências e palestras não

publicadas [...]" (Aquino, 2020, p. 18). A título exemplificativo, o professor cita os seguintes: Lume (UFRGS); Águia (USP); Oasis (Instituto Brasileiro de Ciência e Tecnologia/Ibict); Catálogo de Teses e Dissertações da Capes. Em Portugal, destaca-se o RUN, Repositório da Universidade Nova de Lisboa, além da editora internacional Elsevier.

Sobre a necessária colaboração entre pesquisadores de distintas áreas do conhecimento, Liu (2009) afirma que, muito embora cada pesquisador possa usar ferramentas ou sistemas de forma individual, como processador de texto, o trabalho demanda uma grande equipe de pesquisadores com "habilidades diversas em programação, design de banco de dados, visualização, análise e codificação de texto, estatísticas, análise do discurso, design do site, ética (incluindo assuntos humanos complexos e regras de pesquisa)".[2] Assim, surgem "projetos digitais ambiciosos em nível competitivo com a premissa de fazer a diferença no mundo de hoje. Humanistas trabalhando em colaboração com engenheiros e cientistas sociais [...]"[3] (Liu, 2009, p. 27, tradução nossa).

Já Kaplan (2015, p. 1) aborda três domínios da pesquisa nas humanidades digitais com uso de *big data* a partir de seus respectivos desafios. O primeiro deles prioriza o processamento e a interpretação de grandes conjuntos de dados culturais; o segundo pode ser estruturado em torno das diferentes relações que ligam esses grandes conjuntos de dados, os discursos coletivos, os atores globais e o ambiente informacional; e, por fim, no terceiro domínio, que diz respeito à experiência mais "imersiva" com o *big data*, há o desafio de se lidar com um espaço contínuo contendo as interfaces possíveis organizadas ao derredor desses três polos: imersão, abstração e linguagem. Sem dúvida, ao mesmo tempo em que nos arriscamos no uso de aplicações por nós ainda desconhecidas, as enxergamos também como uma janela de oportunidades capaz de tornar todas as fases da investigação mais íntegras, acessíveis e abertas. Consequentemente, esse acaba sendo o nosso objetivo primordial: tornar a ciência mais integrativa e relevante para a sociedade.

O autor retoma a ideia de humanidades digitais muito mais como comunidade de práticas do que como campo de investigação acadêmica institucionalizado (Kaplan, 2015, p. 2). Tal ideia é similar àquela desenvolvida por Alves, a

2 No original: "it requires a full team of researchers with diverse skills in programming, database design, visualization, text-analysis and -encoding, statistics, discourse analysis, web-site design, ethics (including complex 'human subjects' research rules) etc.".

3 No original: "ambitious digital projects at a grant-competitive level premised on making a difference in today's world. Humanists working on collaborative teams with engineers and social scientists [...]".

qual ele comenta numa entrevista concedida em 2020 (Aquino, 2020, p. 3-4). Todavia, Kaplan avança com a proposta de organizar as HD de *big data* como um campo estruturado, o qual seja caracterizado por metodologias e objetos comuns (em geral, considerando sempre a existência de um grande volume de informações).

A hipótese de trabalho exemplificada busca analisar fluxos e contrafluxos do ecossistema noticioso constituído nessa realidade midiática, onde se inferem, em nível empírico e hipotético, a superestimação de temas periféricos e a disseminação de significados, bem como suas respectivas contribuições para os conhecimentos sobre as temáticas tratadas nas notícias, sejam quais forem. O modo como o jornalismo aborda certos fenômenos tem implicações e consequências na forma como a própria sociedade enxerga tais questões. Obviamente, não se trata de uma via única, mas da fonte de informação cujos resquícios de credibilidade remontam da inquestionável hegemonia conquistada na era moderna, arraigada na *expertise* do jornalismo profissional como instituição.

Segundo Almeida (2014, p. 191), "as conexões entre cultura e tecnologia se tornam cada vez mais estreitas, e não podem ser analisadas de forma ingênua". Partindo da perspectiva mais crítica, e ao retomar revisão sobre a origem e os significados do polissêmico termo "mediação", o autor questiona a cristalizada concepção de que ações de mediação não seriam "simples reação entre dois termos de mesmo nível, mas que em si mesmas seriam produtoras de algo mais"; isso é, agregariam "valor aos processos culturais, informacionais ou comunicacionais, gerando ganhos em termos de conhecimento aos sujeitos envolvidos". Tal concepção se coaduna com a ideia do jornalismo como ação mediadora em sentido amplo, capaz de agregar valores nos três processos.

Conforme elucida Pimenta (2016), a cultura informacional da atualidade, auxiliada pelos dispositivos inerentes à "hipermodernidade", é a forma como realizamos a nossa relação com os aspectos da "visibilidade, da memória, da informação e do conhecimento, [...] potencializados pela possibilidade da mediação, via recursos tecnológicos e suas possibilidades de acesso, convergência e circulação da informação; e pela desintermediação" (Almeida, 2014, p. 192-193 apud Pimenta, 2016, p. 28-29). Ao que parece, estamos imersos num "fenômeno sem precedentes de convergência nestes meios. Dados diferentes, de bases heterogêneas encontram por mediações tecnológicas possibilidades de produzir informações intercruzadas" (p. 22).

> Comunicar e informar de maneira polissêmica, via computação, aplicativos *mobile*, visualizações de massas de dados, indexação, reconhecimento imagético, de

> gráficos e grafias, além de georreferenciamento. Esses são alguns dos desdobramentos capazes de ser usados em um projeto de humanidades digitais. Todos impressões e expressões características de uma cultura informacional tecnológica marcada pelo fenômeno do digital em diferentes espaços e formas. (Pimenta, 2016, p. 25-26)

É nessa complexa ambiência que o jornalimo se reinventa. Ainda que disponha de muitas formas de elaborar/reelaborar discursos, ao se fechar num fazer pretensamente "desinteressado" e simplista, acaba aprofundando preconceitos e repisando lugares-comuns sem debater temas como altruísmo, direitos humanos, sociobiodiversidade e protagonismo dos atores regionais. Debruçarnos cientificamente sobre essa questão permite avaliar, sobretudo, o fazer do jornalismo local e regional; avançando ao propor uma qualificação do discurso midiático na abordagem de temas contemporâneos, em última análise. Não é tarefa fácil, muito embora nos pareça importante de ser aprofundada em face do contexto social, político, econômico e cultural vivenciado hoje.

Nesse ponto, retomamos a lucidez com que Bauman relaciona a democracia (suas qualidades intrínsecas e até sua existência) e a qualidade do *conhecimento*, este que nos pode alcançar de variadas formas e pelo serviço de muitos mediadores, a exemplo do jornalismo:

> A democracia sempre esteve vinculada à qualidade do conhecimento: da *polis* grega ao Iluminismo europeu para o valor mais recentemente atribuído à educação, ao jornalismo investigativo e à opinião pública, a maioria das concepções de democracia reside no sentido de que as pessoas são capazes de pensar e fazer julgamentos por si mesmas.[4] (Bauman et al., 2014, p. 137, tradução nossa)

De fato, a noção de discursos – quanto à elaboração narrativa da notícia – vem ao encontro da perspectiva de autocrítica adotada na pesquisa em comunicação. Todavia, deparamo-nos com a inevitabilidade de tratar um grande volume de dados e com a necessidade de empregar, numa ou em diversas etapas da pesquisa, aplicações capazes de: 1) resgatar publicações, organizando-as conforme variáveis temporais, de origem etc.; 2) sistematizar esses dados segundo os aspectos técnicos, narrativos etc.; 3) analisar o conteúdo e o discurso em

4 No original: "Democracy has always been tied to the quality of knowledge within a demos: From the Greek polis to the European Enlightenment to the value more recently placed on education, investigative journalism and public opinion, most conceptions of democracy rest on some sense that people are able to think and make judgments for themselves".

consonância com o escopo e os objetivos traçados no plano de pesquisa (consideradas as alterações no percurso); e 4) apresentar os resultados de modo a bem representar as principais conclusões alcançadas.

Para analisar os fenômenos intrínsecos à abordagem jornalística em discursos presentes na web, é necessário considerar que o trabalho investigativo se efetiva em torno de um relativamente amplo recorte temporal e da análise de grande volume de dados em sites de notícias previamente elencados. A depender da pesquisa, esse recorte pode ser readequado,[5] todavia, importa saber que existem softwares aptos a realizar tarefas que, pelos padrões de levantamento mediante aplicação de técnicas manuais, seriam avaliadas como muito trabalhosas ou impraticáveis diante da análise de elevados (e diversificados) volumes de material, a exemplo dos mineradores de dados.[6]

A seguir, apresentamos as principais aplicações usadas nas pesquisas em humanidades e ciências sociais hoje para extrair, sistematizar, analisar e visualizar dados em pesquisa social com resgate de grande volume de dados da web, expondo as potencialidades e as limitações de cada uma, a partir da disponibilidade para o acesso/levantamento on-line e gratuito.

Percurso metodológico: delimitação do corpus e organização do conjunto de aplicações

Demo (1996, p. 34) define a pesquisa como sendo o "questionamento sistemático crítico e criativo, mais a intervenção competente na realidade, ou diálogo crítico permanente com a realidade em sentido teórico e prático". Assim entendemos, e por tal razão nos propomos a fazer o reconhecimento inicial – situado como pesquisa documental (Moreira, 2009) – de um conjunto de aplicações que podem auxiliar em nossa investigação principal apresentada no tópico anterior.

Quanto à abordagem, trata-se de pesquisa qualitativa, a fim de considerar tanto conceitos quanto os principais usos das aplicações pelo emprego de técnicas de *webometria*[7] etc. Cumpre salientar: o universo desta pesquisa contempla

5 Poole (2017) alerta que há ferramentas, além daquelas que geralmente usamos (e-mail, Office etc.), como softwares de Análise de Redes Sociais (ARS), de Sistemas de Informações Geográficas (SIGs), de linguagens de programação voltadas para análise de dados, capazes de incorporar conceitos como *big data* e outros.

6 O principal elemento da mineração de dados é a coleção de documentos, haja vista constituírem um conjunto daquele que é o elemento básico colecionável. Como exemplo desse elemento, é possível citar uma "unidade discreta de dados em formato textual", como uma página na web ou um e-mail (Sabino; Heinzle, 2015).

7 Para Barros e Duarte (2009), os trabalhos de mineração de textos estão nos campos: bibliometria,

os sites do Laboratório em Rede de Humanidades Digitais[8] (Larhud), do Digital Humanities Lab (DH Lab)[9] e do colaborativo SourceForge.[10] A escolha dos dois laboratórios se justifica, primeiro, pelo fato de ambos estarem vinculados a instituições tradicionais de pesquisa e desenvolvimento na área das HD. O Larhud é vinculado ao Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), enquanto o DH Lab ao Instituto de História Contemporânea (IHC) da Universidade Nova de Lisboa (Portugal). Em segundo lugar, justifica-se pelo nível de acessibilidade das informações ali contidas (para leigos e usuários de nível básico), além da *expertise* ao aplicar algumas ferramentas em seus projetos institucionais, de modo a testar as funcionalidades e atualizar os dados a respeito das aplicações.

Já o SourceForge aglutina grande variedade de aplicações, permitindo que se alcance um *corpus* significativo de opções para análise. Assim, a primeira fase do processo de seleção dos softwares deu-se com o levamento de aplicações constantes nos dois laboratórios. Em seguida, já tendo definido as quatro etapas de uma pesquisa que considera o *corpus* composto por páginas de notícias na web (coleta, sistematização, análise e visualização de resultados), foram realizadas buscas diretas com essas *keywords* no site SourceForge. Por fim, foram selecionadas aplicações gratuitas que coincidiram com aquelas apresentadas nos dois sites anteriores e outras cujas descrições de uso eram similares àquelas já constantes na listagem prévia.

A fim de compor o conjunto analítico da pesquisa (Fragoso et al., 2011), a lista de aplicações foi tratada segundo essas etapas: 1) fixação do *conjunto* de aplicações e alocação de acordo com a categoria/fase de uso (coleta ou extração, sistematização, análise e visualização de resultados), de acordo com as categorias de exaustividade, representatividade, homogeneidade e pertinência apresentadas por Bardin (2011); 2) elaboração de um quadro composto por nome da ferramenta, principais empregos, título e *link* de acesso (site do desenvolvedor); e, por fim, 3) sistematização de conteúdos e discussão sobre as possibilidades de cada aplicação contemplar a necessidade de pesquisas usando o resgate de páginas da web em sites de notícias ou investigações congêneres.

De modo geral, este artigo se propõe a apresentar um conjunto de aplicações capazes de serem usadas na etapa de tratamento metodológico de pesquisas com

cientometria, infometria, mediametria, museometria e webometria. Tratam aspectos da informação e sua qualidade.

8 Disponível em: http://www.larhud.ibict.br/. Acesso em: 8 jan. 2021.

9 Disponível em: https://dhlab.fcsh.unl.pt/. Acesso em: 8 jan. 2021.

10 Plataforma que compila softwares empresariais e de código aberto com acesso mundial por mais de 32 milhões de usuários. Disponível em: https://sourceforge.net/directory/development/development/os:windows/. Acesso em: 8 jan. 2021.

resgate de grande volume de dados disponíveis na web, a exemplo de notícias de portais. Em particular, a proposta objetiva: 1) elaborar um quadro comparativo entre os softwares utilizados em cada uma das fases; e 2) indicar, no conjunto obtido, quais das aplicações seriam as mais eficazes (custo-benefício, usabilidade, facilidade de gerar *outputs* a outras aplicações sem necessidade de retrabalho, perdas e inconsistências) para investigações que ambicionam resgatar páginas da web (especificamente aquelas de notícias). Importa frisar, no entanto, que a apresentação dos dados é compatível com a disponibilidade e a gratuidade do acesso às informações sobre os softwares, outra razão pela qual se optou por um universo que contempla dois laboratórios institucionais e um site colaborativo.

A seguir, apresenta-se uma disposição das ferramentas retiradas das páginas do Laboratório em Rede de Humanidade Digitais, do Digital Humanities Lab e do site SourceForge. Em resumo, por se entender a delimitação imposta a este texto, o Quadro 1 sintetiza os conteúdos coletados:

Quadro 1 – Seleção de 23 aplicações disponíveis em páginas de HD e de compilados de softwares

Etapa	Aplicação	Finalidade	Disponibilidade
Coleta ou extração de dados	ThoutReader (*last update*: 2016) Indicação: SourceForge	Sistema de ajuda de plataforma cruzada e multidocumento que permite aos usuários navegar, pesquisar, marcar e anexar documentação empacotada em um formato XML extensível. Ele foi escrito em Java e pode ser executado em plataforma que suporte o Java 1.4	Acesso: gratuito (doações) https://sourceforge.net/projects/thout/files/latest/download
	Egonet (*last update*: 2017) Indicação: SourceForge	Programa de coleta e análise de dados de redes egocêntricas. Ajuda a criar o questionário, coletar dados e fornecer medidas gerais de rede global e matrizes de dados para serem usadas para análise posterior por outro software	Acesso: gratuito http://github.com/egonet/egonet
	Web Crawlers (gênero) Indicação: Larhud, HD Lab e SourceForge	Também conhecido como *spider* ou *bot*, é um robô usado pelos buscadores para encontrar e indexar páginas de um site. Ele captura informações das páginas e cadastra os links achados, possibilitando encontrar outras páginas e mantendo sua base de dados atualizada. É empregado para obter dados de páginas web. A tarefa pode ser otimizada pelo *sistema.xml* e pelo *robot.txt*. Exemplos:[11] Arquive-It (serviço *on demand* para construir, gerenciar e pesquisar arquivos), Crawljax (analisa e indexa aplicativos dinâmicos baseados em Ajax), DataparkSearch (*open source*, organiza buscas em site, intranet e sistema local), Googlebot (do Google), Msnbot (pertence ao Bing, mecanismo de busca da Microsoft), Oncrawl (realiza funções de crawler e faz auditorias de SEO nos sites analisados), Wget (oferece coleta e arquivamento remoto de páginas), Yahoo! Sluro	Acesso: gratuito ou comercial *open source* ou não Principais aplicações: Arquive-It Crawljax DataparkSearch Googlebot Msnbot Oncrawl Wget Yahoo! Sluro

11 Lista sintetizada a partir do site comercial https://neilpatel.com/br/blog/web-crawler/. Acesso em: 11 mar. 2022.

Coleta ou extração de dados	Scrapy Indicação: SourceForge	Estrutura de *web scraping* e rastreamento da web rápida e de alto nível. É uma estrutura rápida, de código aberto e de alto nível para *rastrear sites e extrair deles dados estruturados*. Portátil e escrito em Python, pode rodar em Windows, Linux, macOS e BSD. É rápido, simples e extensível, sendo possível escrever as regras para extrair dados e adicionar funcionalidades sem alterar o núcleo, podendo rodar em vários aplicativos. Pode ser usado para mineração de dados, monitoramento e testes automatizados	Acesso: gratuito *open source* https://scrapy.org/ Observação: necessita de assinatura para acesso a funções mais elaboradas para extração
Sistematização de dados	MyWebSQL (*last update*: 2016) Indicação: SourceForge	Cliente de banco de dados WYSIWYG baseado na web escrito em PHP. Possui interface intuitiva com a aparência de um aplicativo de *desktop* e oferece recursos avançados para gerenciamento de banco de dados. Nenhuma instalação é necessária. Basta fazer o *download*, extrair e usar. Ele tem uma versão compacta de arquivo único com funcionalidade total, que pode ser implementada rapidamente no servidor. Atualmente, suporta trabalhar com bancos de dados MySQL, PostgreSQL e SQLite	Acesso: gratuito https://sourceforge.net/projects/mywebsql/files/latest/download Observação: aceita doação para os desenvolvedores
	OmniDB (*last update*: 2020) Indicação: SourceForge	Ferramenta web colaborativa de código aberto para gerenciar banco de dados com foco em interatividade e facilidade de uso pelo design amigável. Baseado em navegador, pode ser acessado em qualquer plataforma. Tem interface responsiva de página única, espaço de trabalho unificado, editor de SQL e outros recursos. Ele suporta Windows, Linux e OSX, e os seguintes SGBDs:[12] PostgreSQL, MariaDB, Oracle, MySQL, SQLite (WIP), Firebird (WIP), SQL Server (WIP), IBM DB2 (WIP)	Acesso: gratuito *open source* https://sourceforge.net/projects/omnidb.mirror/files/latest/download
	Logstash (*last update*: 2021) Indicação: SourceForge	Canal de processamento que recebe dados dinamicamente de várias fontes, os transforma e envia para um arquivo, independentemente do formato ou da complexidade. Ele suporta e ingere dados de todas as formas, tamanhos e fontes, transforma e prepara dinamicamente esses dados e os transporta para a saída à escolha. É extensível, com mais de duzentos *plug-ins* para permitir a configuração de *pipeline*	Acesso: gratuito *open source* https://sourceforge.net/projects/logstash.mirror/files/latest/download
	OpenRefine (*last update*: 2020) Indicação: SourceForge	Escrito em Java e projetado para trabalhar com dados confusos e melhorá-los, sendo possível carregar dados, entendê-los, limpá-los, transformá-los, reconciliá-los e aumentá-los com serviços da web e dados externos, pelo navegador web. Mantém os dados com segurança no computador, executando um pequeno servidor nele, usando o navegador da web para interagir. Permite o compartilhamento externo apenas quando o compilado estiver pronto. Roda em mais de 15 idiomas, é multiplataforma e faz parte do Code for Science & Society[13]	Acesso: gratuito *open source* https://sourceforge.net/projects/openrefine.mirror/files/latest/download
Análise	Voyant Tools Indicação: Larhud	Permite usar textos ou coleções de textos (on-line ou não) para executar funções básicas de mineração de textos. Os produtos gerados – listas de frequência de palavras, gráficos de distribuição de frequência e exibições de KWIC[14] – permitem a extração rápida das características de determinado *corpus* teórico, ampliando a possibilidade de se descobrir temas	Acesso: gratuito *open source* https://voyant-tools.org/

12 Sigla para designar Sistema Gerenciador de Banco(s) de Dados.

13 A organização desenvolve recursos *open source*, apoia a comunidade de tecnologia de interesse público e defende a infraestrutura aberta. Também facilita a colaboração entre organizações e fornece suporte estratégico para iniciativas voltadas para a comunidade dessa área. Disponível em: https://codeforscience.org/. Acesso: 13 dez. 2020.

14 Sigla do termo em inglês *keyword in context*. Trata-se de um processo de pesquisa/classificação de documentos pelo uso de *tags* contadas no texto do referido documento.

Análise	AWStats (*last update*: 2020) Indicação: SourceForge	Analisador de arquivo de *log* de servidor que mostra estatísticas web/mail/FTP, incluindo visitas, visitantes únicos, páginas, acessos, sistema operacional, navegadores, mecanismos de pesquisa, palavras-chave, visitas de robôs, links quebrados etc.	Acesso: gratuito *open source* https://sourceforge.net/projects/awstats/
	Cortext Indicação: Larhud	Ferramenta usada para a análise de um conjuntos de textos com o fim de capacitar pesquisas e estudos abertos em humanidades sobre a dinâmica da ciência, a tecnologia, a inovação e a produção de conhecimento	Acesso: gratuito https://www.cortext.net/
	Sobek Indicação: Larhud	Ferramenta de mineração de texto criada para apoiar aplicações educacionais. Ela tem sido utilizada em várias tarefas, tais como o auxílio aos professores no processo de avaliação de atividades de produção textual, ou de leitura e escrita	Acesso: gratuito http://sobek.ufrgs.br/sobekonline/index.html
	Orange Indicação: Larhud	Possibilita a mineração de dados através de seu software de código aberto que trabalha com *machine learning* e *data visualization*. Seus fluxos de trabalho são baseados em análise de dados interativos atrelados a uma grande opção de ferramentas, incluindo-se aí técnicas de visualização, exploração, pré-processamento e modelagem	Acesso: gratuito *open source* https://github.com/Orange-OpenSource
	RQDA Indicação: Larhud	Pacote R para análise de dados qualitativos assistidos por computador ou CAQDAS. É executado no software estatístico R, mas possui janela separada executando uma interface gráfica do usuário (através do RGtk2). Ele permite uma forte integração da abordagem construtivista da pesquisa qualitativa com a análise quantitativa dos dados, o que aumenta o rigor, a transparência e a validade da pesquisa quali	Acesso: gratuito http://rqda.r-forge.r-project.org/
	Text Ripper Indicação: Larhud	Gera arquivos de textos a partir de PDF ou de páginas web (URL). É uma iniciativa *wiki* capaz de criar métodos de análise on-line	Acesso: gratuito https://wiki.digitalmethods.net/Dmi/ToolTextRipper
	Projeto Lemur (*last update*: 2020) Indicação: SourceForge	Possui busca e mineração de dados e conjuntos de dados ClueWeb. Contempla pesquisa, ferramentas de navegador, ferramentas de análise de texto e recursos de dados que apoiam a pesquisa e o desenvolvimento de *software de recuperação de informação e mineração de texto*, incluindo pesquisa Indri em C ++, *framework* de pesquisa Java, aprendizado RankLib para a biblioteca, conjuntos de dados ClueWeb09 e ClueWeb12 e mineração de dados Sifaka	Acesso: gratuito *open source* https://www.lemurproject.org/ e https://sourceforge.net/projects/lemur/files/latest/download
Visualização de resultados	Raw Graphs Indicação: SourceForge	Estrutura de visualização de dados de código aberto capaz de tornar a representação visual de dados complexos mais simples. Criado para designers e *geeks*, RAW Graphs busca fornecer um elo perdido entre aplicativos de planilha (por exemplo, Microsoft Excel, Apple Numbers, OpenRefine) e editores de gráficos vetoriais (por exemplo, Adobe Illustrator, Inkscape, Sketch)	Acesso: gratuito *open source* https://rawgraphs.io/about
	Google Data Studio Indicação: SourceForge	Ferramenta que transforma dados em relatórios e painéis informativos, fáceis de ler, de compartilhar e de personalizar. Gera os relatórios a partir de fontes de dados tratadas conforme os objetivos do usuário	Acesso: gratuito https://www.google.com/
	Gephi Indicação: Larhud, DH Lab	Software de visualização e exploração para todos os tipos de gráficos e redes. Ele funciona em Windows, Mac OS X e Linux e tem o objetivo de ajudar analistas de dados ao permitir o uso de interfaces interativas e de *outputs* que facilitam a visualização	Acesso: gratuito *open source* https://gephi.org/

Visualização de resultados	Grafana (*last update*: 2021) Indicação: SourceForge	Plataforma de análise e monitoramento projetada para todos os bancos de dados. Ela permite visualizar e entender métricas por meio de painéis dinâmicos e reutilizáveis baseados em dados. Permite criar e compartilhar, além de ampla exploração de métricas e registros, e de alertas sobre as métricas mais importantes	Acesso: gratuito *open source* https://grafana.com/ Observação: aceita contribuição
	Nuvens de palavras (gênero) Indicação: Larhud, HD Lab e SourceForge	Há aplicativos gratuitos que oferecem a possibilidade de criação de nuvens de palavras a partir de qualquer texto, de modo fácil e intuitivo, como Tagul, TagCrowd, Word it out, Tag Cloud Generator, ou todos aqueles sugeridos no blog 21st Century Educational Technology and Learning.15	Acesso: gratuito/comercial http://www.larhud.ibict.br/nuvem-de-palavras/
	Tulip (*last update*: 2021) Indicação: SourceForge	Estrutura de visualização de informação dedicada à análise e visualização de dados relacionais. O Tulip fornece uma biblioteca completa, apoiando o design de visualização interativa de informações (Université de Bourdeux, França)	Acesso: gratuito https://tulip.labri.fr/site/ e https://sourceforge.net/projects/auber/files/latest
	Timeline JS Indicação: Larhud	Aplicação para criar linhas do tempo em temas e períodos determinados pelo usuário. É alimentada por um *dataset*[16] e usa tecnologia do KnightLab, este vinculado à Northwestern University, nos EUA	Acesso: gratuito *open source* https://timeline.knightlab.com/

Fonte: compilado a partir de levantamentos em Larhud, DH Lab e SourceForge (2020, on-line).

Possibilidades de aplicação em pesquisas com resgate de páginas da web (notícias)

Nesse ponto, a proposta é avançar na discussão mais qualificada a partir do compilado de aplicações representadas no Quadro 1. Para tanto, o primeiro passo foi selecionar aquelas que seriam mais adequadas aos propósitos expostos no segundo tópico deste artigo, aqui retomados: 1) coleta (ou extração) de dados; 2) sistematização de dados; 3) análise (conforme os tipos de dados e os *outputs* da fase antecedente, por exemplo, CSV); e, por fim, 4) visualização de resultados. Conforme já explicitado, tomou-se por pressuposto que uma investigação na qual se trabalhe com o resgate de páginas da web (notícias publicadas durante certo período de tempo em determinados sites) pode ser dividida em etapas nas quais é potencialmente possível usar uma ou mais dessas aplicações.

15 Fonte secundária da pesquisa (não acessada diretamente). Disponível em: https://21centuryedtech.wordpress.com/.

16 *Dataset* é o conjunto de valores de entrada (planilha) usados pelo software para a criação da linha do tempo.

Em síntese, a escolha das aplicações descritas, a despeito do universo de opções, pautou-se nos seguintes quesitos: custo-benefício (disponibilidade gratuita dos softwares e de todos os seus recursos) e adequação das funcionalidades às necessidades metodológicas inerentes ao escopo de pesquisa considerado. Especificamente em relação a esse tópico, no qual se restringe ainda mais a seleção de softwares, um fator decisivo foi a disponibilidade da documentação e dos tutoriais de tais aplicações na web, permitindo sua operacionalização inclusive por usuários iniciantes.

Ao se referirem à web como "os arquivos da vida contemporânea", Brugger e Finnemann assim explicam:

> a) O arquivo da web é um arquivo em tempo real (*real-time*), pois necessita de uma atitude urgente em relação ao que se quer preservar, enquanto ainda está on-line; b) O arquivo da web não é uma cópia de tudo o que estava on-line, mas uma nova versão, "renascida" e única, o que significa definir o que arquivar, o que omitir, como tratar as atualizações dos sites, programas e estratégias de arquivamento e disponibilização. Isso implica ser reativo às mudanças tecnológicas durante algum tempo, buscando atualizar constantemente as metodologias de arquivamento para preencher essa lacuna; c) O arquivo da web torna-se multitemporal e multiespacial, primeiro porque possui diversas versões de determinado site, e pode ser navegado por data de captura de conteúdo. Multiespacial, pois, dependendo dos períodos capturados, poderá haver mais ou menos partes arquivadas, tornando a extensão diferente ao longo do tempo. (Brugger; Finnemann, 2013 apud Rockembach, 2019, p. 134-135)

Ao articular as humanidades digitais e o arquivamento de conteúdos disponíveis na web, Rockembach (2019, p. 131) ressalta a relevância de usar "políticas, metodologias e tecnologias que envolvem a seleção, captura, armazenamento, preservação e disponibilização de conteúdo da web para acesso e uso retrospectivo", fator que pode ser reforçado pela aplicação do senso de comunidade oriundo das HD na tarefa de preservar a memória digital para permitir o seu acesso e uso futuro. A ideia é possibilitar que as informações sejam recuperadas de modo eficaz, auxiliando no resgate de conteúdos para a realização de pesquisas comparativas ou analíticas, por exemplo.

Coleta ou extração de dados

Do conjunto de aplicações listadas e passíveis de uso na fase de coleta/extração de dados, é interesante perceber o quanto cada uma delas se vincula a tarefas

bastante específicas. Nesse sentido, avalia-se como mais adequado para as pesquisas com resgate de notícias o Scrapy, capaz de rastrear sites e extrair dali dados estruturados. Além de ser gratuito, extensível e *open source*, é escrito em Python e tem a vantagem de rodar em vários sistemas operacionais. Multiplataforma, ele também permite que o usuário defina regras de extração sem a necessidade de modificar os comandos gerais. Ele é empregado por pessoas físicas e jurídicas, inclusive para tratar notícias. A empresa Parsely, por exemplo, usa o Scrapy para copiar artigos de centenas de sites noticiosos.

Antes, contudo, é preciso entender seu uso para resgatar conteúdos "brutos", como páginas de períodos de tempo relativamente longos, acessando muitos dados. Assim, não sendo possível resgatá-los pela busca simples nos portais de notícias que compõem o universo, é indispensável o uso de ferramentas de busca remota na web, isto é, no arquivo digital de sites.[17] Ora, o jornalismo é em tudo dependente dos meios digitais, mas também das estruturas que o sustentam como uma instituição, sendo assertivas as palavras de André Lemos:

> Retire do "jornalismo" a internet, as empresas jornalísticas, universidades e professores de jornalismo, os jornaleiros, os distribuidores, os computadores, os celulares, os órgãos reguladores, o papel jornal, a web... e veja se você ainda vê algum "sujeito" livre de amarras! [...] A genialidade e originalidade de uma ação não vêm da independência de outros actantes, mas justamente do contrário: das boas associações estabelecidas. (Lemos, 2011, p. 18-19 apud Saad Corrêa; Carlan da Silveira, 2017, p. 169)

Com efeito, tais autoras resumem de modo brilhante o contexto atual e os atores envolvidos nas experiências de consumo do jornalismo dito ubíquo. Segundo elas, a produção jornalística da atualidade "baseia-se numa rede de agentes mediadores – humanos e não humanos – do processo comunicacional, enquanto peças capazes de interferir nas características de uma experiência de consumo de informação". Assim, assumindo que "a rede digital – por onde transitam notícias, informações e dados – está presente e imperceptível em todos os lugares e espaços [... e que] ela é ubíqua" (Saad Corrêa; Carlan da Silveira, 2017, p. 166), o jornalimo também o é. O desafio, ao tomá-lo como objeto e organizá-lo como *corpus* a fim de extrair significados, não deve passar incólume diante das acuradas lentes das humanidades digitais.

17 Para esse fim, consultar: https://archive.org/web/ ou https://arquivo.pt/. Acesso em: 23 nov. 2020.

Sistematização de dados (e preparação)

Em se tratando da sistematização, essa é a etapa na qual o conjunto de dados passa por uma padronização capaz de adequá-lo ao próximo passo da pesquisa, que é propriamente a análise. O que se espera dessa aplicação é a eficiência para organizar e categorizar os dados, preparando-os conforme as necessidades de *inputs* do software ou módulo analítico, elaborando tabelas, listas de conteúdos separados por vírgula, distribuindo elementos por tipo de mídia e em bancos de dados, além de "empacotar" os dados em formatos e extensões mais comuns e "usáveis". No Quadro 1 estão listadas quatro aplicações de acesso gratuito, interessante vantagem nesse tipo de software.

Desse modo, outro aspecto foi determinante para a escolha de duas aplicações mais condizentes com a delimitação do objeto empírico deste trabalho: Logstash e OpenRefine. Ambas têm códigos de fonte aberta e atualizações recorrentes, o que confere confiabilidade na proteção contra *bugs* e outros erros de execução. O Logstash pode receber dados dinâmicos de várias fontes e tem como *output* um arquivo compilado que independe da complexidade dos *inputs*, além de ter mais de duzentos *plug-ins* de saída. Já o OpenRefine é multiplataforma e baseado em Java (mais acessível), propondo tornar dados confusos de *input* em informações inteligíveis e compartilháveis. Assim, crê-se na viabilidade do uso de tais aplicações na fase de sistematização nesse tipo de pesquisa.

Análise de dados (grande volume/heterogêneos/multimídia)

Relativamente ao terceiro momento delimitado e passível de ser implementado com a ajuda de aplicações, precisamente a etapa de análise, o desafio é obter software, sistema, ferramenta ou módulo capaz de extrair significado de dados heterogêneros, multimidiáticos e em grande volume. Conforme explanado, a tarefa resume-se a receber um compilado de dados previamente organizados, muito embora complexos: no caso por nós teorizado, as notícias veiculadas em sites, resgatadas para posterior organização e análise. Assim, tratando-se de pesquisa com uma coleção de notícias, elas eventualmente serão compostas por conteúdos de diferentes formatos, fato esse que agregaria complexidade ao conjunto de dados submetido à análise pelo software selecionado.

No Quadro 1, há oito aplicações que podem ser usadas em investigações do gênero, todas elas de acesso não comercial, a saber: Voyant Tools, AWStats, Cortext, Sobek, Orange, RQDA, Text Ripper e Projeto Lemur. Com a vantagem de serem todas gratuitas, a aplicações podem ser *open source*, algo que

representaria uma vantagem para os usuários em nível avançado. Ademais, observou-se – mediante descrições de funcionalidade nas páginas dos desenvolvedores – que boa parte delas têm funções básicas similares a aplicações pagas,[18] a exemplo da mineração de textos.

Diante do exposto, priorizou-se, como mais provável escolha de uso nessa etapa de análise metodológica, a mais eficaz para a geração de significados ao tratar grandes volumes de dados em vários formatos: Orange. Esse software privilegia o aprendizado de máquina e a visualização de dados para uma análise de aspectos interativos, ou seja, seu uso é bastante dinâmico.

Visualização de dados (representação de resultados)

Finalmente, após selecionar, organizar, analisar e interpretar o *corpus* investigado, parte-se para uma importante etapa, qual seja a visualização de resultados obtidos pós-análise. Frisa-se o interesse, em alinhamento com a concepção de ciência aberta, de disponibilizar *outputs* de todas as etapas já elencadas, inclusive para possibilitar o reúso do compilado noutras investigações ou permitir que os pares possam referendar os métodos empregados e até questioná-los/criticá-los, o que gera *insigths* e faz a ciência progredir. Há muitas aplicações aptas a tornar os resultados de pesquisas mais claros, acessíveis e até autoexplicativos.

Acerca do acesso aberto, Poole (2017) destaca a tensão existente entre o modelo proposto pela academia (valorizando a difusão irrestrita dos conhecimentos produzidos) e a resistência do mercado editorial, um ponto de vista a partir do qual se alega a perda de receita das editoras, o incentivo ao plágio e o enfraquecimento dos direitos autorais. Essas questões merecem o devido aprofundamento – que extrapola o presente debate, a despeito de tais desafios reclamarem formas inovadoras de enfrentamento, como uma necessária adaptação da revisão de pares em trabalhos nativos da web e dos modelos de *copyright* nesse ambiente (Poole, 2017, p. 12).

Na seleção do artigo, são listados sete programas para a visualização de dados de pesquisas, sendo todos eles distribuídos de forma gratuita, *open source* ou não. No presente trabalho, é de se destacar a necessidade de que os *inputs*

18 Embora o levantamento tenha se concentrado nos softwares gratuitos, mencionam-se dois dos sistemas comerciais largamente empregados por investigadores da área de ciências humanas com vistas à análise de grandes volumes de dados extraídos da web: Atlas.ti e NVivo. Ambas as ferramentas comerciais têm como vantagem a disponibilidade de módulos complementares que podem ser usados em outras fases do tratamento de dados, desde a extração até a visualização e o compartilhamento, com a desvantagem de seus custos serem calculados em dólar.

compreendam uma gama de extensões, de modo que não haja retrabalho, inconsistências ou perda de dados entre as etapas de análise e visualização, por exemplo, mas não somente aí. Ilustramos: é perfeitamente possível criar nuvem de palavras com dados extraídos ainda na fase de coleta, ou gerar relatórios interativos após a sistematização. Por fim, a propriedade de compartilhamento multimídia e multiplataforma também é um diferencial.

Considerando o exposto, entende-se que estas aplicações poderiam auxiliar na visualização de dados conforme a delimitação empírica teorizada no artigo: Raw Graphs, Grafana e Timeline JS. A primeira é *open source*, de distribuição gratuita, e cria representações visuais com dados complexos, sendo sua maior vantagem o acesso a planilhas (por exemplo, OpenRefine) e editores gráficos vetoriais (por exemplo, Ilustrator). Já a Grafana é uma plataforma *open source* de análise e monitoramento, que permite visualizar métricas e registros em painéis dinâmicos e reutilizáveis. O Timeline JS possibilita a criação de linhas do tempo dinâmicas e interativas a partir de um *dataset*. Também é uma ferramenta *open source* e está disponível para *download* gratuito no site do KnigthLab, da Northwestern University (EUA). Além dessas aplicações, os geradores de nuvens de palavras são acessíveis e representam visualmente os dados de tipo textual (exemplos: Tagul, TagCrowd etc.), sendo plenamente passíveis de ser incorporados às estratégias de visualização ao longo da pesquisa.

Considerações

Neste texto, nos propusemos a expor uma compreensão inicial e organizada de um universo de programas, módulos e ferramentas capazes de serem acionados em investigações no contexto das humanidades digitais. O desafio foi pensar de que maneiras essas aplicações poderiam auxiliar e até conduzir a resultados e *insights* no desenvolvimento de pesquisas com resgate de páginas da web, considerando-se as metodologias e a epistemologia próprias das HD, hoje utilizadas com relativa timidez. Felizmente, mostra-se promissor o seu emprego na área da comunicação, em particular nos estudos que visem ao tratramento de um relativamente grande volume de notícias da web ou de coleções de dados heterogêneos e multimidiáticos de natureza congênere.

Em linhas gerais, as tarefas propostas se resumiram a apenas duas. Primeiro, foram listadas as aplicações possivelmente capazes de atender às necessidades de investigação quanto à coleta, sistematização, análise e visualização de dados. Em seguida, estabeleceu-se uma abordagem mais qualitativa na avaliação, segundo critérios preestabelecidos, das vantagens das aplicações e das funções

que podem realizar, considerado o objeto teorizado para este artigo. Em suma, há muitos projetos colaborativos e *open source* acessados por pesquisadores e profissionais das ciências humanas e sociais e boa parte deles são vinculados a instituições de ensino/pesquisa na Europa e nos Estados Unidos.

Nesse cenário, a principal desvantagem das iniciativas é a limitação tanto das ferramentas quanto dos tutoriais e demais documentos à língua inglesa (algumas delas ao alemão ou francês), tornando-as mais acessíveis a quem tem um domínio pelo menos básico do idioma. Em segundo lugar, muito em decorrência de uma separação canônica entre áreas científicas, a falta de literacia digital (e midiática)[19] limita o acesso e a operacionalização de ferramentas que poderiam ser úteis na elucidação de questões metodológicas quando do tratamento de grande volume de dados.

Ainda que não seja o propósito desta pesquisa, é importante mencionar que boa parte dos pesquisadores em ciências sociais utiliza largamente os softwares Atlas.ti[20] e NVivo[21] em suas investigações, questão que poderia ser objeto de novos levantamentos comparativos entre aplicações gratuitas e comerciais. Em geral, as interfaces amigáveis dos módulos pagos acabam por defini-los como opções hegemônicas entre os cientistas das humanidades. Todavia, existem ferramentas gratuitas que, associadas ou individualmente, seriam plenamente capazes de entregar resultados satisfatórios. O Orange¸ por exemplo, relativamente à etapa de análise do tratamento metodológico, privilegia aspectos interativos em coleções de dados de tipos variados.

Em última análise, é profícua e louvável a prática interdisciplinar pelos pesquisadores das ciências humanas e sociais no ambiente digital, de onde muitos de nós extraem seus problemas de pesquisa, mas também boa parte dos recursos capazes de ajudar a compreender essas questões.

19 Muitos estudos têm abordado o tema da *media literacy*, inclusive para conhecer práticas e competências relevantes e até indispensáveis, classificando-as. Apronfundar sobre a temática em Lopes (2015).

20 Contempla ferramentas analíticas que permitem a visualização interpretativa do conteúdo, além da capacidade de analisar documentos primários (textos, imagens, vídeos etc.) e em grande volume.

21 Busca informações em dados não estruturados da web (por exemplo, notícias), tendo como objetivo realizar análises em níveis profundos nos diferentes volumes de dados.

Referências

ALMEIDA, Marco Antônio de. Mediação e mediadores nos fluxos tecnoculturais contemporâneos. *Informação e Informação*, v. 19, n. 2, 2014. Disponível em: http://www.uel.br/revistas/uel/index.php/informacao/article/view/20000/pdf_24. Acesso em: 7 jan. 2021.

AQUINO, Israel. Digital humanities como uma comunidade de práticas: entrevista com o professor Daniel Alves (IHC/NOVA FCSH). *Revista Aedos*, v. 12, n. 26, p. 740-761, 2020. Disponível em: https://seer.ufrgs.br/aedos/article/view/105132. Acesso em: 11 jan. 2021.

BARDIN, Laurence. *Análise de conteúdo*. São Paulo: Edições 70, 2011.

BARROS, A.; DUARTE, J. *Métodos e técnicas de pesquisa em Comunicação*. São Paulo: Atlas, 2009.

BAUMAN, Zygmunt et al. After Snowden: Rethinking the impact of surveillance. *International Political Sociology*, v. 8, n. 2, p. 121-144, 2014. Disponível em: https://doi.org/10.1111/ips.12048. Acesso em: 20 dez. 2020.

DEMO, Pedro. *Pesquisa e construção de conhecimento*. Rio de Janeiro: Tempo Brasileiro, 1996.

DHLAB. *Digital Humanities Lab*. Disponível em: https://dhlab.fcsh.unl.pt. Acesso em: 23 nov. 2020.

FRAGOSO, Suely; RECUERO, Raquel; AMARAL, Adriana. *Métodos de pesquisa para internet*. Porto Alegre: Sulina, 2011.

KAPLAN, Frédéric. A map for big data research in digital humanities. *Frontiers in digital humanities*, v. 2, 2015. Disponível em: https://www.frontiersin.org/articles/10.3389/fdigh.2015.00001/full. Acesso em: 15 nov. 2020.

LARHUD. *Laboratório em Rede de Humanidades Digitais* [*Ferramentas em HD*]. Disponível em: http://www.larhud.ibict.br/index.php?title=-Ferramentas. Acesso em: 23 nov. 2020.

LEWIS, S. C.; WESTLUND, O. Big data and journalism: Epistemology, expertise, economics and ethics. *Digital Journalism*, v. 3, issue 3, p. 447-466, 2015. Disponível em: http://dx.doi.org/10.1080/21670811.2014.976418. Acesso em: 7 out. 2020.

LOPES, Paula. Avaliação de competências de literacia mediática: instrumentos de recolha de informação e opções teóricometodológicas. *Media e Jornalismo*: educação para os media na era digital, n. 27, v. 15, n. 2, 2015. Coimbra University Press: 2015. Disponível em: https://digitalis-dsp.uc.pt/handle/10316.2/38139. Acesso em: 16 mar. 2022.

LIU, Alan. Digital humanities and academic change. *English Language Notes*, v. 47, n. 1, p. 17-35, 2009. Disponível em: https://alanyliu.org/wp-content/uploads/2018/06/dh-and-academic-change-page-proofs.pdf. Acesso em: 16 nov. 2020.

MAURI, M.; ELLI, T.; CAVIGLIA, G.; UBOLDI, G.; AZZI, M. RAWGraphs: A visualisation platform to create open outputs. Biannual Conference on Italian Sigchi Chapter, 12., 2017, New York. *Proceeding*, New York, USA: ACM, 2017. p. 1-5. Disponível em: https://doi.org/10.1145/3125571.3125585. Acesso em: 11 jan. 2021.

MOREIRA, S. V. Análise documental como método e como técnica. In: DUARTE, J.; BARROS, A. *Métodos e técnicas de pesquisa em comunicação*. São Paulo: Atlas, 2009.

PATEL, Neil. *Web Crawler*: entenda o que é, quando usar e como funciona. Disponível em: https://neilpatel.com/br/blog/web-crawler/. Acesso em: 11 mar. 2022.

PIMENTA, Ricardo M. Os objetos técnicos e seus papeis no horizonte das humanidades digitais: um caso para a ciência da informação. *Revista Conhecimento e Ação*, v. 1, n. 2, 2016. Disponível em: https://revistas.ufrj. br/index.php/rca/article/view/20. Acesso em: 12 jan. 2021.

POOLE, A. H. The conceptual ecology of digital humanities. *Journal of Documentation*, v. 73, n. 1, 2017. Disponível em: https://www.researchgate.net/publication/311921946_The_Conceptual_Ecology_of_Digital_ Humanities. Acesso em: 23 jan. 2021.

ROCKEMBACH, Moisés. Arquivamento da web no contexto das humanidades digitais: da produção a preservação da informação digital. *Liinc em Revista*, Rio de Janeiro, v. 15, n. 1, p. 131-139, 2019. Disponível em: http://revista.ibict.br/liinc/article/view/4578/4142. Acesso em: 2 dez. 2020.

SAAD CORRÊA, Elizabeth; CARLAN DA SILVEIRA, Stefanie. Proposta teórico-metodológica para a pesquisa de objetos no jornalismo. *Matrizes*, São Paulo, v. 11, n. 2, p. 163-182, maio/ago. 2017. Disponível em: https://www.revistas.usp.br/matrizes/article/download/133850/133228/. Acesso em: 13 jan. 2021.

SABINO, Allan Renato; HEINZLE, Roberto. Ferramenta para construção de ontologia a partir de dados não estruturados. In: COMPUTER ON THE BEACH, 2015, Itajaí. *Anais*... Itajaí: Univali, 2015. Disponível em em: https://siaiap32.univali.br/seer/index.php/acotb/article/view/7029. Acesso em: 12 dez. 2020.

SOURCEFORGE. *The complete open-source and business software platform create, collaborate & distribute to over 32 million users worldwide*. Disponível em: https://sourceforge.net/. Acesso em: 8 jan. 2021.

ZENG, M. L. Smart data for digital humanities. *Data and Information Science*, v. 2, n. 1, p. 1-12, 2017. Disponível em: https://swib.org/swib19/slides/01_zeng-keynote.pdf. Acesso em: 17 dez. 2020.

Recebido em 28/5/2021
Aprovado em 10/3/2022