DIGITALIZAÇÃO DE JORNAIS

UMA REFLEXÃO SOBRE DESAFIOS E MELHORES PRÁTICAS

NEWSPAPERS DIGITIZATION

A REFLECTION ON CHALLENGES AND BEST PRACTICES

Bruno Leal Pastor de Carvalho | Doutor em História Social pelo PPGHIS/UFRJ; mestre em Memória Social pelo PPGMS/Unirio; professor substituto de Teoria e Filosofia da História do Departamento de História da UFF; professor-tutor do curso EAD de História da Unirio

RESUMO

O avanço das mídias digitais tem permitido a digitalização de jornais e sua disponibilização na internet, desafiando as instituições a repensarem seus acervos. A prática historiográfica também tem sido provocada a estabelecer novas reflexões interdisciplinares. Este artigo tem o objetivo de pensar criticamente a prática de digitalização de jornais, procurando destacar desafios, limites, análise de projetos e melhores práticas no campo.

Palavras-chave: digitalização; jornais; fontes históricas; fontes digitais.

ABSTRACT

The advance of digital media has allowed the digitization of newspapers, available on the internet, challenging the institutions to rethink their historical collections. The historiographical practice has also been led to establish new deep interdisciplinary reflections. This article intends to examine the newspaper digitization practice, seeking to highlight challenges, limits, to analyze projects and to review best practices in this field.

Keywords: digitization; newspaper; historical sources; digital sources.

RESUMEN

El avance de los medios digitales ha permitido la digitalización de periódicos y su disposición en internet, desafiante las instituciones a replantear sus colecciones. La práctica historiográfica también ha sido provocada al fin de establecer nuevas reflexiones interdisciplinares. Este artículo tiene como objetivo pensar críticamente la práctica de digitalización de periódicos, tratando de poner de relieve los desafíos, los límites, el análisis de los proyectos y las mejores prácticas en el campo.

Palabras clave: digitalización; periódicos; fuentes históricas; fuentes digitales.

INTRODUÇÃO

Manuel Castells tem explicado o mundo contemporâneo a partir de uma revolução tecnológica concentrada nas tecnologias da informação (Castells, 2007, p. 39). Essa revolução, aponta o autor, tem sido responsável por modelar em ritmo acelerado a base material da sociedade e engloba, atualmente, não só a internet, mas também os dispositivos móveis de comunicação, as redes sociais e todo tipo de mídia digital. Este artigo discute uma faceta desta transformação: a crescente digitalização de jornais. O trabalho está dividido em quatro partes. Na primeira, apresento brevemente o fenômeno da digitalização. Na segunda, explico porque é importante digitalizar. Na terceira, exploro alguns projetos e práticas com os quais podemos aprender. Na quarta e última, debruço-me sobre um caso específico – o *Jornal do Brasil* – que nos permite compreender como dois processos de digitalização podem ter resultados completamente diferentes. A fim de evitar perda de sentido, optei por reproduzir neste artigo alguns termos técnicos em inglês, uma vez que não existe uma tabela de equivalência para o português. Essa reflexão, por fim, embora traga a perspectiva de um historiador, diz respeito a outros profissionais que também lidam diretamente com a história.

A DIGITALIZAÇÃO DE JORNAIS

Poucas fontes históricas permitem tantas possibilidades de investigação para o historiador quanto o jornal. Estou me referindo a duas dimensões elementares que esse meio de comunicação comporta. Primeiro, a sua dimensão discursiva, isto é, a sua habilidade para ordenar o mundo, estabelecer fatos, produzir consenso e emprestar sentido à experiência histórica. Depois, mas não menos importante, a sua capacidade para registrar os mais distintos fenômenos culturais, políticos, econômicos, sociais e até mesmo naturais. Segundo Wilhelm Bauer, "o jornal é uma verdadeira mina de conhecimento: fonte de sua própria história e das situações mais diversas; meio de expressão de ideias e depósito de cultura. Nele encontramos dados sobre a sociedade, seus usos e costumes, informes sobre questões econômicas e políticas" (Bauer, 1970, p. 85).[1] Não fortuitamente, os jornais – e a imprensa em geral – estão cada vez mais presentes nas pesquisas historiográficas. Segundo levantamento de Ana Paula Goulart, "do total de trabalhos que abarcam o século XX apresentados, em 1995, no Encontro Nacional de Pós-Graduandos em História, cerca de 70% utilizavam meios de comunicação (sobretudo jornais) como fonte histórica" (Ribeiro, 1999, p. 1).

Nos últimos anos, a importância dos jornais para os estudos acadêmicos se tornou ainda maior graças à digitalização de um grande número de títulos, dos jornais pequenos e locais aos grandes e de circulação internacional, dos não correntes àqueles que ainda encontram-se nas ruas. A partir dos anos 2000, os historiadores se viram diante de um universo quase

1 Importante sublinhar, no entanto, que os jornais não são meros depositários do real ou geradores de um discurso neutro. O discurso jornalístico está interessado na elaboração simbólica deste real.

inesgotável de reportagens, notícias, editoriais, cartas de leitores, anúncios, notas, colunas sociais e crônicas. Ao mesmo tempo, bibliotecas, museus, arquivos e universidades se viram deparados com a necessidade de estabelecer critérios para o tratamento desse material. Para se ter uma ideia da transformação em curso, o microfilme, que era até então o suporte mais recente em termos de preservação de documentos escritos, surgiu ainda no século XIX (Pinheiro; Moura, 2015).

O *The New York Times*, um dos jornais mais influentes do mundo, digitalizou todo o seu acervo histórico (1851 até o presente) recentemente. Segundo seus cálculos, mais de 13 milhões de artigos estão disponíveis na Internet.[2] Os britânicos *The Guardian* e *The Observer*, controlados pelo mesmo grupo, somaram esforços e fizeram o mesmo. Seu conteúdo vai do final do século XVIII aos primeiros anos da década de 2000.[3] No Brasil, os principais jornais seguiram a mesma linha: *O Globo*,[4] *O Estado de S.Paulo*[5] e *Folha de S.Paulo*[6] digitalizaram todo o seu acervo histórico. Para além dos veículos de maior poder financeiro, muitos dos "pequenos" também foram digitalizados por projetos coletivos. Um deles é o *Newspaper Archive*, iniciativa americana que possui mais de dois bilhões de artigos, agregando jornais de 23 países, publicados de 1607 aos dias atuais.[7] Podemos mencionar ainda o *Periodika*,[8] da Letônia, a Biblioteca Digital Hispânica,[9] da Espanha, e a Hemeroteca Digital,[10] no Brasil.

POR QUE DIGITALIZAR?

Há muitas razões para digitalizar acervos históricos. Três me parecem fundamentais. Em primeiro lugar, a digitalização ajuda a democratizar o acesso ao conhecimento. Durante muito tempo, arquivos, museus e bibliotecas foram os fiéis e exclusivos depositários dos documentos. Seu acesso era limitado: era preciso ir pessoalmente a esses espaços físicos – geralmente localizados em grandes capitais – para consultar aquilo que é patrimônio público. E mesmo quando se tinha acesso a essas instituições, era preciso contar com a disponibilidade do documento, que podia já estar sendo consultado por outro usuário ou ausente para higienização ou restauro, por exemplo. A lógica da Internet tem ajudado a subverter esta lógica. Uma vez digitalizado e disponibilizado na Web, o acervo dessas instituições se torna de fato público. Os ganhos que derivam daí dizem respeito não só ao historiador, mas a toda a sociedade. É o caso de documentos que o trabalhador deve recuperar para obter

2 Conferir: <http://www.nytimes.com/ref/membercenter/nytarchive.html>.
3 Conferir: <https://www.theguardian.com/info/2012/jul/25/digital-archive-notice>.
4 Conferir: <http://acervo.oglobo.globo.com>.
5 Conferir: <http://acervo.estadao.com.br>.
6 Conferir: <http://acervo.folha.uol.com.br>.
7 Conferir: <http://newspaperarchive.com>.
8 Conferir: <http://periodika.lv>.
9 Conferir: <http://www.bne.es/es/Catalogos/HemerotecaDigital>.
10 Conferir: <http://memoria.bn.br/>.

benefícios sociais ou do indivíduo que no passado foi perseguido por uma ditadura e agora precisa reunir evidências a fim de ser indenizado pelo Estado. O exercício da cidadania e a garantia da Justiça são, desta maneira, questões influenciadas pela digitalização de acervos ou pelo acesso ágil e facilitado a documentos e informações públicas.

Em segundo lugar, temos a questão da preservação. Qualquer documento corre o risco do desaparecimento. O desgaste começa com o próprio manuseio do original por parte do pesquisador, mesmo que sejam adotados rigorosos procedimentos de consulta. A exposição à luz, à umidade e à própria manipulação do pesquisador são fatores que comprometem a integridade do documento. Além disso, não devemos nos esquecer dos casos de roubos, incêndios, alagamentos, depredação e mau acondicionamento, que podem levar à deterioração irreversível ou à destruição completa. Mofo, fungos ou bactérias, danosos não só aos documentos, mas aos que os manipulam, também devem ser considerados ameaças. Com as cópias digitais, sobretudo em tempos de redes sociais *on-line*, o desaparecimento de uma fonte histórica torna-se bastante improvável. O caso do Museu da Língua Portuguesa, em São Paulo, ilustra muito bem os potenciais da digitalização. Em dezembro de 2015, o Museu foi completamente destruído por um incêndio que, além de fazer uma vítima fatal, um bombeiro que trabalhava no local, queimou completamente o seu patrimônio material. A tragédia para seu acervo histórico, no entanto, foi minimizada, pois boa parte estava preservada em servidores e discos rígidos guardados em outros lugares.[11]

Em terceiro lugar, a gestão da informação. Este é, talvez, o aspecto que mais tenha revolucionado a prática historiográfica. Uma vez que os documentos são transpostos para o meio digital, toda a informação contida neles pode ser indexada. Podemos definir indexação como "um arranjo sistemático de entradas desenhado para permitir que usuários localizem informações em um documento" (Harrison; Wyman, 2006, p. 37).

A indexação é uma prática antiga, preexistente ao digital, sendo tradicionalmente feita por bibliotecários, museólogos e arquivistas, haja vista que todo acervo precisa ser catalogado e sistematizado. Com o advento do digital, a indexação passou a ser realizada, também, por sofisticados *softwares* capazes de ler uma grande massa de informações e torná-los "encontráveis" de forma muito mais rápida. A indexação está no cerne dos buscadores de informação. Com o auxílio de softwares de indexação, o historiador ganha tempo. Isso se torna evidente quando pensamos, por exemplo, na pesquisa em jornais e revistas. Há alguns anos, os pesquisadores precisavam dedicar semanas e meses a procura de determinado termo ou acontecimento. Graças à indexação, isso pode ser feito em segundos. Quando inserimos uma palavra-chave em um campo de busca e "damos *enter*", acionamos uma complexa cadeia de cálculos que vai percorrer todo o documento e nos oferecer resultados bastante confiáveis. Dispondo de mais tempo, o pesquisador pode se dedicar

11 Conferir esta informação em "Museu da Língua Portuguesa tem 'backup' do conteúdo, diz curadora", no G1, em 21 de dezembro de 2015. Disponível em: http://g1.globo.com/sao-paulo/noticia/2015/12/museu-da-lingua-portuguesa-tem-arquivo-de-todo-conteudo-diz-curadora.html. Acesso: 10 mar 2016.

mais a leituras, a análises, à própria escrita e à ampliação do escopo da pesquisa. Em 1994, já ciente do poder do meio digital, Barbara Quinte sublinhou: "The ocean flows of on-line information are all streaming together, and the access tools are becoming absolutely critical. If you don't index it, it doesn't exist. It's out there but you can't find it, so it might as well not be there"[12] (Zafran, 1998, p. 30).

DIGITALIZAÇÃO DE JORNAIS: MELHORES PRÁTICAS E PROJETOS DE SUCESSO

Sendo a digitalização de jornais uma realidade muito recente, as instituições encarregadas de salvaguardar e difundir a memória ainda estão amadurecendo aquilo que podemos chamar de "melhores práticas". É um processo que leva tempo, pois exige investimento financeiro, demanda profissionais de um novo tipo e *feedback* dos usuários. Porém, o acúmulo de importantes experiências, nos últimos anos, já tem nos permitido trabalhar com alguns parâmetros e direções no campo da digitalização de jornais. Nesse sentido, vale a pena refletir sobre o caso norte-americano.

Quando os primeiros jornais começaram a ser digitalizados nos Estados Unidos, entre o final da década de 1990 e meados da década seguinte, não faltavam opções tecnológicas no mercado. Algumas eram de ótima qualidade, enquanto outras apresentavam resultados bem pouco satisfatórios. Em comum, praticamente todas pertenciam às empresas que as comercializavam, isto é, não podiam ser transferidas integralmente para o comprador. Após alguns anos, esse tipo de parceria se revelou danosa. Os jornais digitalizados eram atrelados a sistemas de marca registrada. Isso significava gastos constantes para as empresas contratantes do serviço. Bibliotecas, arquivos, museus e universidades viram-se presas a um modelo dispendioso e pouco flexível. Eram ainda impedidas de migrar para outros formatos e, em alguns casos, chegavam, até mesmo, a perder o controle sobre os seus próprios conteúdos (University of California, 2011, p. 1).

Em 2005, esse panorama mudou. O National Endowment for the Humanities (NEH) e a Library of Congress (LC) criaram o *National Digital Newspaper Program* (NDNP), que tinha duas metas: (1) criar uma plataforma digital que reunisse jornais de todos os estados americanos publicados no período 1836-1922 e (2) criar um portfólio de melhores práticas para a digitalização de jornais. (UC, 2011, p.1). As duas metas foram alcançadas. A primeira, por meio da criação do *Chronicling America*, um diretório *on-line* de jornais americanos publicados no período mencionado (1836-1922). Atualmente, este diretório supera dez milhões de páginas digitalizadas. Já a segunda meta foi alcançada através de diversas publicações com a rubrica de Technical Guidelines & Specifications, que estabeleceram etapas básicas a serem seguidas por projetos de digitalização: (I) *inventorying*;

12 Em livre tradução: "Os fluxos do oceano de informações *on-line* estão todos fluindo juntos e as ferramentas de acesso estão se tornando absolutamente críticas. O que não é indexado, não existe. Se está por aí, mas não se pode encontrar; então pode muito bem não estar em lugar algum".

(II) *organizing*; (III) *Format Management*; (IV) *Metadata Packaging*; (V) *Checksum Management*; (VI) *Packaging* (Skinner & Shultz, 2014, p.11). Quanto ao escaneamento em si, o NDNP estipula::

> Digital reproductions should be made from a preservation copy of microfilm, a clean second-generation duplicate silver negative; Technical scanning requirements: maximum resolution possible between 300-400 dpi, relative to physical dimensions of the original material; 8-bit grayscale – TIFF 6.0 uncompressed; Two-up film should be split so that there is one page image per file; De-skew images that contain text blocks exhibiting skew of greater than 3 degrees (Greater skew leads to less accurate OCR); Crop to include visible edge of page, retaining up to ¼ inch beyond edge; Optional: Capture microfilm target frames. These image files will be identified in the reel metadata but will not be used for display. Capture additional scanning resolution targets, i.e., 35mm Grayscale Preservation Microfilm Target, (2 images per reel--target as specified by LC) at the start of each session, to monitor scan quality. These scan target images should be delivered with microfilm target images and page images and identified in reel metadata (Library of Congress, 2011, p. 5-6).

Para falar mais sobre os critérios desenvolvidos pela NDNP nada melhor do que examinar os projetos de digitalização que se apoiam neles. É o caso do *The Center for Bibliographical Studies and Research* (CBSR), da Universidade da Califórnia, que lançou o *California Newspaper Collection* (CDNC), em 2006. Hoje, o projeto possui mais de 450 mil páginas de jornais digitalizadas daquele estado americano. O CDNC segue várias orientações técnicas e metodológicas inspiradas nos manuais da NDNP: avaliar a qualidade e quantidade do conteúdo disponível, determinar quais títulos são mais valiosos (tendo em vista a limitação dos fundos disponíveis), privilegiar jornais publicados antes de 1923 (por conta dos direitos autorais), procurar escanear jornais microfilmados em detrimento do material impresso (a digitalização a partir de microfilmes é mais eficiente e barata), produzir cópias de segurança, produzir inventários, metadados e assegurar que sejam usados formatos digitais compatíveis (University of California, 2011, p. 2-5).

Outro projeto bastante influenciado pelo NDNP é o *Utah Digital Newspapers* (UDN), fruto de uma parceria firmada entre a University of Utah, a Brigham Young University e a Utah State University com o intuito de digitalizar e disponibilizar na Internet, gratuitamente, jornais históricos do estado de Utah. O projeto foi lançado ainda em 2002, mas só ganhou impulso com o apoio financeiro e técnico da NDNP. Em 2007, a UDN já tinha mais de 570 mil páginas de jornais digitalizadas em seus servidores. De 60 visitas por dia em 2003, o *site* do projeto pulou para 830 em 2006. A iniciativa ganhou diversos prêmios, tais como o *Award of Merit*, da American Association for State and Local History, e o *John Award*, da Utah Press Association (Herbert; Estlund, 2007). Entre as diretrizes da NDNP utilizadas pelo UDN estão: a utilização de mapas para apontar a origem geográfica dos jornais reunidos na base de dados; a indexação de todo o conteúdo; a colaboração dos

usuários; a utilização dos microfilmes para a digitalização; a formação de uma equipe interdisciplinar; a indicação do contexto histórico, entre outras, conforme podemos ler a seguir:

> The entire collection is full-text searchable through a search box on the home page. […] One of the best and most popular ways to browse for regional news is our county map. […] Using the advanced search feature, users may limit searches to a particular group of titles and specific fields. […] The more practical and technical considerations are availability, quality, and format of the source materials. From our own experience, we determined that the image quality scanned from original paper sources was superior to those from microfilm. […] When microfilm is used, the master microfilm reels must be located and copied, because patron copies being used for research are often too scratched and worn. […] Once we locate good source materials, the UDN advisory board reviews the titles. Our board is an extremely knowledgeable group of local historians, librarians, writers, and industry representatives. […] The board also provides additional historical context which cannot be found in a normal catalog record. […] We also consider user requests, once we add a portion of a county paper, local demand surges for more content from that county to be added either from a rival paper or the current title online. […] Each page goes through an article "zoning" process where human beings identify and classify them as news, an advertisement, or birth, death, or marriage announcement. An automated process performs Optical Character Recognition (OCR) against each article and creates a file of the article text. After generating the "raw" text, another automated process filters it through English dictionaries, a Utah place-names dictionary, and an extensive surnames list. The OCR-generated text is not 100 percent accurate. (It averages 70 percent, according to our own survey.) Still, it provides keyword access to the content that is impossible with microfilm. Two people separately transcribe the masthead and article headlines and subheadings. This insures that headings and subheadings are nearly 100 percent accurate (Herbert; Estlund, 2007, p. 338-340).

É importante notar que nos Estados Unidos, diferente do que ocorre em outros países, como o Brasil, muitos projetos de digitalização de jornais são empreendimentos quase exclusivos de bibliotecas universitárias. As universidades – sejam públicas ou privadas – contam com diversos financiamentos, inclusive do governo federal, que lhes permitem digitalizar dezenas de jornais locais. Para isso, contam com equipamentos, laboratórios, servidores e profissionais capacitados. Esses projetos servem não apenas aos pesquisadores e estudantes da universidade, mas à sociedade de uma forma geral, uma vez que o material é disponibilizado gratuitamente na internet. O modelo é aplicado com sucesso graças ao sistema de *endowment* de muitas universidades. O *endowment* "consiste na criação de um patrimônio perpétuo que gera recursos contínuos para a conservação, expansão e promoção de uma

determinada atividade, por meio da utilização dos rendimentos desse patrimônio".[13] O fundo é formado por doações feitas por organizações e antigos alunos, e pode ser usado de forma desburocratizada para diversos fins: construção de alojamentos para estudantes, criação de institutos, montagem de laboratórios ou financiamento de projetos de digitalização.

Bastante consolidadas, as referências do NDNP extrapolaram as fronteiras norte-americanas. Em 2009, elas foram adotadas, por exemplo, no processo de digitalização de vários jornais árabes, a maioria publicada na virada do século XIX para o XX, depositados na Biblioteca da Mesquita Al-Aqsa, na Jerusalém oriental, e cujo acesso, até aquele momento, era restrito basicamente às autoridades palestinas municipais. O projeto tinha o objetivo de preservar a rara coleção e expandir o seu acesso. Foram selecionados 24 títulos, entre revistas e jornais. Tendo em vista os manuais do NDNP, realizou-se: escaneamento em alta resolução (300 dpi), arquivos de baixa compressão, uso de formatos não proprietários e uso do *Optical Character Recognition* (OCR) para converter as imagens escaneadas em texto. O projeto obteve êxito, mas não deixou de encontrar desafios, que são parte integrante da experiência de digitalização:

> The project faced a number of challenges due to external factors as well as those directly related to undertaking a large digitization project of historical newspapers. The quality of the original papers, including different text characters, irregular fonts, text density, torn or smudged pages, and a variety in layout posed many challenges during the image capture process. The project team also realized that the digitization process for long-term preservation is very challenging and time-consuming, taking much longer to scan and create digital master files than originally expected. In addition, the project faced a shortage of trained staff and significant budget shortfalls because of the global economic downturn (Matusiak; Harb, 2009, p. 9).

No Brasil, o Conselho Nacional de Arquivos (Conarq) tem publicado importantes documentos de referência, como a *Carta para preservação do patrimônio arquivístico digital* (2005),[14] *Recomendações para digitalização de documentos arquivísticos permanentes* (2010),[15] *Diretrizes para a presunção de autenticidade de documentos arquivísticos digitais* (2012)[16] e *Diretrizes para a implementação de repositórios arquivísticos digitais confiáveis* (2015).[17] Porém, o debate sobre digitalização tem sido marginal no meio acadêmico brasileiro. Podemos conjecturar uma série de motivos para explicar esse cenário. Os departamentos de história, biblioteconomia, ciência da informação e arquivologia dialogam muito pouco. Há pouca si-

13 Conferir: <http://edireitogv.com.br/o-endowment/como-funciona>.

14 Conferir: <conarq.gov.br/images/publicacoes_textos/Carta_preservacao.pdf>.

15 Conferir: <conarq.gov.br/images/publicacoes_textos/Recomendacoes_digitalizacao_completa.pdf>.

16 Conferir: <conarq.gov.br/images/publicacoes_textos/conarq_presuncao_autenticidade_completa.pdf>.

17 Conferir: <conarq.gov.br/images/publicacoes_textos/diretrizes_rdc_arq.pdf>.

nergia entre os professores e estudantes desses cursos, assim como é enxuto o número de publicações e conferências feitas conjuntamente. Também vamos notar que, no Brasil, os grandes projetos de digitalização, com exceção da hemeroteca digital da Biblioteca Nacional, são conduzidos pelos próprios veículos de comunicação. Soma-se a isso, ainda, os custos altos da digitalização e as dificuldades relativas à negociação dos direitos autorais.

A despeito do diminuto debate, o país acumula algumas experiências importantes. A seguir, vamos discutir uma delas: a digitalização do *Jornal do Brasil*, um dos mais importantes veículos da história da imprensa brasileira. Fundado em 1891, no Rio de Janeiro, por Rodolfo Dantas, o *Jornal do Brasil* destaca-se por sua longevidade (1891-2010, impresso; 2010 até o presente, *on-line*), por seu peso político e pela reformulação estética e editorial que influenciou a imprensa ao longo das décadas de 1950 e 1960 (Ribeiro, 2007). O acervo histórico do *Jornal do Brasil* foi contemplado com dois projetos de digitalização. Seus resultados, no entanto, foram bastante diferentes.

O JORNAL DO BRASIL: DUAS DIGITALIZAÇÕES, DOIS RESULTADOS DISTINTOS

O anúncio da primeira digitalização do *Jornal do Brasil* (JB) foi feito em 2010 e a iniciativa coube ao *Google News Archive*, uma ferramenta criada havia dois anos pelo Google. Na ocasião de seu lançamento, a companhia norte-americana anunciou da seguinte maneira o seu novo empreendimento: "nós estimamos que há bilhões de páginas ao redor do mundo contendo cada história já contada. É nosso objetivo ajudar os leitores a encontrar todas elas, do menor jornal semanal local ao maior jornal diário nacional".[18] Embora ambiciosa, a iniciativa tinha um modelo bastante simples: o Google entrava com a tecnologia e os jornais autorizavam a digitalização de suas edições antigas. Para o Google, as vantagens eram enormes: o *site* atrairia mais acessos, a empresa acumularia ainda mais dados sobre os seus usuários e dominaria mais um segmento de buscas na *Web*. Para os jornais, a parceria chegava no tempo apropriado: justamente quando os editores testemunhavam suas vendas avulsas e de assinaturas caírem, a digitalização do acervo poderia divulgar a marcar e atrair novos leitores.

O projeto realizado pelo *Google News Archive*, no entanto, teve várias limitações. Em primeiro lugar, o acervo do JB não foi digitalizado na íntegra. Não há edições no período de 1891-1929, exceto pelo solitário exemplar de 31 de dezembro de 1910. Entre a década de 1930 e 1990, a maioria das edições foi digitalizada, mas há falhas abundantes em diversos anos. De 2000 a 2010, não se encontra disponível para consulta qualquer edição do jornal. Em segundo lugar – o que soa bastante surpreendente para uma empresa que revolucionou a busca na Internet –, o *Google News Archive* não fez a indexação do conteúdo ou esta não foi oferecida ao usuário. Em outras palavras, não é possível fazer buscas por palavras-chave.

18 Official Google Blog: *Bringing history on-line, one newspaper at a time*. 8 out. 2008. Disponível em: <https://googleblog.blogspot.com.br/2008/09/bringing-history-online-one-newspaper.html>. Acesso: 26 maio 2016.

Dessa forma, resta ao pesquisador explorar o jornal à moda antiga, página por página, o que anula a maior vantagem do material digitalizado.

O *Google News Archives* também não oferece ao usuário qualquer informação sobre os jornais digitalizados. Não sabemos nada sobre os fundadores do JB ou sobre sua trajetória. Não há tutoriais, informações de contato ou botões para serem acionados em caso de descoberta de erros no acervo. A interface, no entanto, é intuitiva. O usuário pode selecionar cinco tipos de organização das edições: dia, semana, mês, ano e década. Uma vez escolhida a opção, o sistema exibe na tela uma composição de miniaturas. Estas, por sua vez, quando clicadas, levam o leitor para "dentro" da edição. As páginas podem ser passadas, uma a uma, como um carrossel. E tanto a ferramenta de ampliação quanto a qualidade da digitalização podem ser consideradas satisfatórias. Ainda assim podemos observar problemas consideráveis: a página do *Google News Archive* mistura inglês e português; na década de 1890, verifica-se erro no processo de digitalização dos rolos do jornal: há apenas duas edições e elas são de 1996; na barra de navegação, quando o usuário folheia o jornal, não é indicado o ano, a edição ou o caderno do jornal. Finalmente, a numeração das folhas apontada no monitor nem sempre coincide com a paginação do veículo.

O Google nunca revelou detalhes sobre a equipe responsável pelo *Google News Archive* ou os trâmites por trás do seu processo de digitalização de jornais. Podemos inferir, no entanto, que o seu programa não levou em consideração alguns dos cuidados básicos que vimos nas páginas anteriores deste artigo: indexação, cobertura completa do acervo, contexto histórico do veículo digitalizado, participação dos usuários no processo de correção de erros, entre outros. Desse modo, o acervo histórico de um dos jornais mais importantes da imprensa brasileira foi levado à Internet de forma precária, oferecendo poucas vantagens para os pesquisadores e o grande público. Não surpreende, portanto, que o projeto *Google News Archive* tenha sido interrompido em nível global, em 2011, embora os jornais que já tinham sido digitalizados não tenham sido retirados do ar. Segundo o que um porta-voz do Google informou na época: "os internautas podem continuar pesquisando os jornais que já estão disponíveis, mas nós não planejamos acrescentar novas ferramentas e funcionalidades ao *Google News Archives*, nem aceitar novos microfilmes ou arquivos digitais para processamento".[19]

Um ano depois da descontinuidade do *Google News Archive*, a Biblioteca Nacional (BN) anunciou uma nova digitalização do JB. Dessa vez, o material foi adicionado à Hemeroteca Digital, maior projeto digital daquela instituição. O JB passou a compor um banco de dados que ultrapassa atualmente 600 títulos brasileiros e estrangeiros digitalizados. São jornais e revistas que vão desde o século XVIII, como o raríssimo *Folheto de Lisboa*, passando por clássicos do século XIX, como *O Jornal das Senhoras* e *O Paiz*, chegando a jornais contempo-

19 O Globo, Google desiste do projeto de digitalização..., 20 jun. 2011. Disponível em: <http://oglobo.globo.com/economia/tecnologia/google-desiste-do-projeto-de-digitalizacao-dos-arquivos-de-jornais-2767399>. Acesso em: 29 fev. 2016.

râneos ao *Jornal do Brasil*, caso do *Correio da Manhã* e do *Diário de Notícias*. A Hemeroteca Digital tem a chancela do Ministério da Cultura e é reconhecida pelo Ministério da Ciência e Tecnologia. Sua construção contou com o apoio da Financiadora de Estudos e Projetos (Finep), que possibilitou a compra de equipamentos e o pagamento de servidores, empresas parceiras e pessoal.

A Hemeroteca Digital, apesar de lançada em 2011, é desdobramento de um projeto mais antigo da BN, lançado em 2006, de inserção da Instituição no mundo digital: a Biblioteca Digital.[20] De acordo com Angela Bittencourt, atual coordenadora da Biblioteca Digital, pouco depois que a Hemeroteca Digital foi lançada, o empresário Nelson Tanure, atual proprietário do JB, entrou em contato com a Biblioteca Nacional a fim de negociar uma nova digitalização do acervo de seu jornal. Tanure estava insatisfeito com o trabalho realizado pelo Google e acreditava que a Hemeroteca Digital poderia fazer melhor. "Quando o projeto surgiu", diz Bittencourt, "nós nem pensávamos na inclusão do Jornal do Brasil. A ideia inicial do projeto era a digitalização e a disponibilização da nossa coleção de jornais que já estavam em domínio público. O JB não estava".[21]

A segunda digitalização do *Jornal do Brasil* levou mais ou menos seis meses para ser concluída. A captura das imagens foi feita a partir dos microfilmes da própria Biblioteca Nacional, conforme recomendado pela NPND. Bittencourt lembra que era digitalizada uma média de 20 mil fotogramas por dia. Devido ao grande volume de páginas, a Biblioteca Nacional terceirizou parte do trabalho. Os microfilmes – que não podiam deixar o prédio da Instituição – eram digitalizados, usando-se *scanners* próprios, mas manipulados pela equipe de uma empresa de tecnologia que já era parceira da Biblioteca Nacional na Hemeroteca Digital, a DocPro, com sede no Rio de Janeiro. Depois de realizado este processo, o material seguia, em vários HDs, para a sede da DocPro, onde os arquivos eram indexados, cortados, montados e revisados. Uma vez concluída esta etapa, o material estava pronto para ser disponibilizado na Internet. Atualmente, a Hemeroteca Digital, vale dizer, usa os seus próprios servidores para armazenar todo o conteúdo de sua base.[22]

Embora as duas versões digitalizadas do JB coexistam hoje na Internet, as diferenças entre seus resultados são enormes, com ampla vantagem para o trabalho brasileiro. Em primeiro lugar, a Biblioteca Nacional digitalizou todas as edições deste jornal carioca. Não há falhas e nem lacunas, a não ser aquelas originais da fonte, do próprio *Jornal do Brasil*. Em segundo lugar, a Hemeroteca Digital é uma plataforma muito mais intuitiva, detalhada e organizada que o *Google News Archive*. Tudo é explicado ao visitante. O projeto da Biblioteca Digital, por exemplo, possui nove subguias: apresentação, políticas de digitalização, missão, histórico, laboratório de digitalização, estatísticas da BN Digital, normas e padrões, parcerias e "quero colaborar". No *site*, também há informações sobre o JB e sua história. Quanto ao material

20 Conferir: <https://bndigital.bn.br/>.

21 Entrevista de Angela Bittencourt ao autor, em 24 de fevereiro de 2016.

22 Idem.

digitalizado, ele apresenta ótima resolução e carrega em alguns poucos segundos, dependendo da conexão e da extensão da pesquisa. Quando comparamos o *Google News Archive* e a Hemeroteca Digital, o primeiro leva vantagem em apenas dois critérios: a visualização do jornal em sistema de carrossel (que permite ter uma melhor noção de cada edição) e a velocidade de carregamento das páginas, que é ligeiramente mais rápida.

Porém, o grande diferencial entre os dois projetos diz respeito à indexação do conteúdo. No Google News Archive, como vimos, a indexação não foi realizada. Na Hemeroteca Digital, todos os jornais são indexados. De acordo com os engenheiros Ernesto Breitinger e José Lavaquial, dois dos cinco sócios da DocPro, responsável pela indexação de todo o acervo da Hemeroteca Digital, a digitalização é a parte menos importante do trabalho que desenvolvem: o foco está na disponibilização da informação.[23]

Na versão da BN, a busca pode ser feita por períodos (divididos em décadas) e/ou por palavra-chave. Também é possível acessar datas específicas e a pesquisa pode englobar mais de um jornal ao mesmo tempo. Em questão de segundos, a procura é realizada pelo sistema e oferecida ao usuário.

Quanto à indexação da Hemeroteca Digital, utiliza-se uma tecnologia exclusiva de Reconhecimento Óptico de Caracteres (OCR) da DocPro, cuja maior virtude está na "aprendizagem". Em um jornal do século XIX, por exemplo, a palavra *locus* pode ter sido escrita como *lovus*. Para que o programa de indexação identifique a palavra como *locus*, a equipe da DocPro pode produzir uma regra de equivalência. É como funcionam os atuais processadores de texto de nossos computadores: quando digitamos uma palavra que não consta no seu dicionário, o *software* a sublinha. Nós, então, podemos adicionar essa palavra e esta será reconhecida da próxima vez que for utilizada. No entanto, no caso da digitalização do acervo, essa depuração é feita pelos engenheiros da empresa e a participação dos usuários talvez seja um ponto a ser considerado. Afinal, a maior virtude também é o maior desafio: uma pesquisa em jornais mais antigos ou com erros na reprodução na Hemeroteca ainda pode apresentar mais erros na detecção de palavras via OCR.

CONSIDERAÇÕES FINAIS

Quando nos referimos à digitalização de jornais, não existe uma fórmula única a ser seguida. Cada caso pode apresentar particularidades de acordo com as características do jornal em questão: o formato físico, a ausência de microfilmes, a língua, o financiamento que o seu depositário possui para a digitalização, a tecnologia empregada etc. Tudo isso faz diferença. No entanto, uma vez que já acumulamos 20 anos de experiência em digitalização de jornais, podemos já estabelecer diretrizes e recomendações básicas que podem ser aplicadas, mesmo sob adaptação, a qualquer projeto no campo. A NDNP estabeleceu várias diretrizes importantes neste sentido. Não há dúvidas, por exemplo, quanto à necessidade

23 Entrevista de José Lavaquel ao autor, 25 de fevereiro de 2016.

de uma equipe interdisciplinar, à indexação do conteúdo, à contextualização histórica do jornal, à flexibilidade da tecnologia utilizada ou à participação dos usuários, fundamentais na qualificação do material digitalizado. Quando alguns desses parâmetros não são seguidos, ou pelo menos não em sua totalidade, existirão projetos com falhas sensíveis, caso da digitalização do JB pelo *Google News Archive*.

Deve-se destacar que a transformação provocada pelas novas mídias, conforme sublinhou Manuel Castells no início deste artigo, não representa para o ofício do historiador apenas mudanças no suporte da fonte. A digitalização de jornais vem inaugurar uma nova maneira de fazer e compreender a pesquisa histórica. Em primeiro lugar, a "pesquisa digital" – como podemos chamar a pesquisa em meios digitais e com objetos digitais – demanda entendimento da linguagem tecnológica e de aspectos técnicos ligados ao campo da comunicação. Como vimos, o historiador precisa saber o que é e como funciona um mecanismo de indexação. Ele precisa estar apto ainda a avaliar diferentes projetos de digitalização, os seus limites e suas possibilidades, sempre tendo em vista que estes projetos podem produzir resultados muito diferentes, mesmo quando são desenvolvidos a partir de um mesmo acervo – e que estes resultados podem impactar diretamente em sua pesquisa.

Em segundo lugar, mas igualmente importante, o historiador precisa desenvolver novas capacidades e competências para pesquisar no universo digital, já que este possui lógica própria. É possível, por exemplo, extrair resultados mais ou menos precisos, dependendo da maneira como se usa um campo de busca por palavra-chave. Dentro de um arquivo digital, também precisamos reconhecer as interconexões entre documentos, estabelecer roteiros de investigação, saber salvar imagens, converter formatos, reconhecer falhas, transpor barreiras técnicas e manipular programas de computadores para sistematizar descobertas. A pesquisa digital não suplanta a "pesquisa tradicional", isto é, aquela feita nos arquivos convencionais, mediante a experiência tátil do documento, mas ocorre em paralelo a esta, o que deve nos parecer natural, uma vez que nosso mundo agora também é digital.

Referências bibliográficas

BAUER, Wilhelm et al. *A imprensa como fonte histórica*. São Paulo: Departamento de Jornalismo e Editoração da ECA-USP, 1970.

CASTELLS, Manuel. *A sociedade em rede*. São Paulo: Paz e Terra, 2007.

HARRISON, Larry; WYMAN, Pilar. Frequently Asked Questions About Indexing. In: ZAFRAN, Enid L. (ed.). *Starting an indexing business*. S.l.: Information Today Inc., 1998.

HERBERT, John; ESTLUND, Karen. Creating Citizen Historians. *Western Historical Quarterly*, Logan, v. 39, n. 3, p. 333-341, 2008.

LIBRARY AND INFORMATION CONGRESS, 2009, Milão. *Conference paper...*, E-prints in Library and Information Science, 2009.

LIBRARY OF CONGRESS (EUA). *The National Digital Newspaper Program* (NDNP) *Technical Guidelines for Applicants*. Washington, 2011. Disponível em: <http://www.loc.gov/ndnp/guidelines/archive/NDNP_201113TechNotes.pdf>. Acesso em: 30 maio 2016.

MATUSIAK, Krystyna; HARB, Qasem Abu. Digitizing the Historical Periodical Collection at the Al-Aqsa Mosque Library in East Jerusalem. In: IFLA WORLD LIBRARY AND INFORMATION CONGRESS, 2009, Milão. *Conference paper...*(S.l.), E-prints in Library and Information Science, 2009.

PINHEIRO, Alejandro de Campos; MOURA, Paloma de Leles de. A microfilmagem. *Múltiplos Olhares em Ciência da Informação*, Belo Horizonte, v. 4, n. 2, 2015, ISSN 2237-6658.

RIBEIRO, Ana Paula Goulart. Mídia e história: ambiguidades e paradoxos. *Eco* – Publicação da Pós-Graduação da Escola de Comunicação da UFRJ, Rio de Janeiro, v. 4, n. 1, p. 5-10, 1999.

_____. *Imprensa e história no Rio de Janeiro dos anos 50*. Rio de Janeiro: E-papers, 2007.

SKINNER, Katherine; SCHULTZ, Matt. *Guidelines for Digital Newspaper Preservation Readiness*. Atlanta: Educopia Institute, 2014. Disponível em: <http://digital.library.unt.edu/ark:/67531/metadc282586/m2/1/high_res_d/Guidelines_for_Digital_Newspaper_Preservation_Readiness.pdf>. Acesso em: 30 maio 2016.

UNIVERSITY OF CALIFORNIA. Center for Bibliographical Studies and Research. *A Guide and Best-Practices for Institutions around the Golden State*. Los Angeles, 2011. Digitizing California's Newspapers Collection. Disponível em: <http://cdnc.ucr.edu>. Acesso em: 30 maio 2016.

ZAFRAN, Enid L. (ed.). *Starting an indexing business*. S.l.: Information Today Inc., 1998.

Recebido em 1/6/2016
Aprovado em 9/8/2016