O acesso aos jornais históricos: Considerações sobre o desenvolvimento de coleções digitalizadas

Access to historical newspapers: Considerations on the development of digitised collections

https://doi.org/10.14195/2183-5462_39_8

Olívia Pestana
Universidade do Porto, Faculdade de Letras
Centro de Investigação Transdisciplinar Cultura, Espaço e Memória - CITCEM
opestana@letras.up.pt

Resumo

A utilização de recursos eletrónicos no desenvolvimento de processos de investigação que envolvam a recolha de dados a partir de jornais históricos tem-se desenvolvido de uma forma crescente, sendo de realçar a digitalização dos jornais. O conhecimento das necessidades informacionais dos investigadores e a utilização de um abrangente conjunto de funcionalidades técnicas contribuem para o melhoramento da conceção de bases de dados de jornais digitalizados e das suas formas de pesquisa. O presente artigo pretende explorar algumas questões determinantes no desenvolvimento e utilização de coleções digitalizadas de jornais históricos, ou seja, abordar aspetos fundamentais como os que se relacionam com as práticas de pesquisa e utilização das coleções, as limitações técnicas decorrentes da digitalização e, ainda, as funcionalidades das interfaces do utilizador e as modalidades de pesquisa que as bases de dados podem oferecer.

Palavras-chave

jornais históricos; jornais digitalizados; interfaces do utilizador; modalidades de pesquisa

Abstract

The use of electronic resources for developing research processes involving the collection of data from historical newspapers has been increasingly developed. In this context, the digitisation of newspapers is of particular relevance. Knowledge of researchers' information needs and the use of a comprehensive set of technical features contribute to improving the design of digitised newspaper databases and their search options. This paper aims to explore some key issues in the development and use of digitised collections of historical newspapers, in other words, it addresses fundamental aspects such as those related to the search practices and use of the collections, technical limitations arising from digitisation, functionalities of user interfaces, and also search modalities that databases can exhibit.

Keywords

historical newspapers; digitised newspapers; user interfaces; search options

Introdução

A utilização de recursos eletrónicos no desenvolvimento de processos de investigação que envolvam a recolha de dados a partir de jornais históricos tem vindo a tornar-se uma necessidade. É de realçar a facilidade de acesso à informação viabilizada pela digitalização de jornais históricos, realizada por organizações públicas ou privadas em todo o mundo, particularmente na última década (Gooding, 2016b, p. 1; Mussell, 2012, p. 28). A digitalização permite ultrapassar as vicissitudes decorrentes da difícil localização dos jornais e respetivas coleções completas, a distância relativamente a acervos físicos e a dificuldade em consultar exemplares em precário estado de conservação.

Efetivamente, e conforme reconhecido por Late e Kumpulainen (2021), a digitalização tem influenciado o trabalho dos historiadores em pelo menos duas formas: por um lado, a maioria das coleções exibe disponibilidade omnipresente, ou seja, os investigadores podem utilizá-las a partir dos seus próprios dispositivos eletrónicos de trabalho, sem quaisquer impedimentos geográficos ou temporais; por outro lado, muitas interfaces de plataformas eletrónicas que servem de acesso a coleções de jornais digitalizados oferecem a possibilidade do uso de outras tecnologias que exploram novos caminhos na interpretação dos textos (Brake, 2012, p. 222–223).

Os modernos instrumentos técnicos de análise de conteúdo aplicáveis aos estudos de texto promoveram a utilização do texto digitalizado, pois, através do recurso a avançadas técnicas de pesquisa e análise do conteúdo, permitem uma análise automatizada ou semi-automatizada de largos volumes de textos, até recentemente impraticável. Esta possibilidade viabiliza a obtenção de resultados diferentes de uma análise manual e perspetivam conclusões que de outro modo não seriam atingíveis.

A utilização de jornais digitalizados revela alguns problemas daí decorrentes, nomeadamente as alterações da leitura integral dos jornais que ilustram opções editoriais e conexão entre notícias, o que ainda pode ser efetuado, mas a possibilidade de pesquisa ao nível da notícia vem trazer novas observações ao longo do texto e pode favorecer uma menor atenção à análise global de cada página e/ou edição. Por outro lado, a leitura decorrente da ordenação cronológica também fica comprometida com o fomento da pesquisa e acesso aos textos das notícias (Hansen e Paul, 2015, pp. 7–8). Estas questões resultam da aplicação das técnicas de pesquisa que se implantaram e desenvolveram com os motores de busca da internet, bem como com as práticas de pesquisa de artigos científicos em bases de dados de literatura científica, em especial de artigos de publicações periódicas (Mussell, 2012, p. 59). Para apoiar o processo de pesquisa de informação, os mais diversos sistemas de informação têm sido desafiados a disponibilizar mecanismos para o acesso, apresentação e exploração de informação, bem como a mostrar estas representações para facilitar a interpretação e ainda a apoiar a extração e manuseio de informação a partir das mais diversas representações do conhecimento (Marchionini, 1995, pp. 140–141).

A cada vez mais frequente prática de pesquisa em linha, o recurso a pontos de acesso normalizados, ou seja, recorrendo a princípios estabelecidos uniformemente de âmbito internacional, são de especial relevância, mas a sua aplicação é desconhecida do investigador em geral, estando o seu uso entregue a bibliotecários ou a estudantes e investigadores devidamente capacitados. Nos pontos de acesso cruciais

à pesquisa em jornais históricos, encontram-se os que se baseiam em sistemas de organização do conhecimento e de gestão de entidades, como *thesauri*, classificações, taxonomias, e, ainda, autoridades de nome (Hitchcock, 2013).

É neste contexto que o presente artigo pretende explorar algumas questões determinantes no desenvolvimento e utilização de coleções digitalizadas de jornais históricos, ou seja, abordar aspetos fundamentais como os que se relacionam com as práticas de pesquisa e utilização das coleções, as limitações técnicas decorrentes da digitalização e os critérios, e propriedades técnicas a considerar.

A utilização das coleções de jornais históricos digitalizados

O desenvolvimento dos processos de digitalização de jornais históricos, bem como o desenho das bases de dados que disponibilizam a pesquisa e o acesso a esses jornais, será verdadeiramente efetivo se corresponder às necessidades informacionais dos investigadores e às suas práticas de interação com as diversas plataformas digitais. É também importante identificar as práticas anteriormente existentes num contexto menos eletrónico, pois, como refere Brake (2012, p. 222), a maioria dos historiadores tem consultado o conteúdo dos jornais em busca de outros tópicos, mas um conjunto de investigadores pesquisou o próprio meio, o que requer o acesso à navegação e consulta dos jornais digitalizados ao nível da página e/ou do número completo, considerando, aqui, o conceito de navegação como uma aproximação ao "folhear" das páginas impressas. E é neste âmbito que se pode encontrar muito do trabalho de digitalização realizado, ou seja, apenas para disponibilizar páginas completas e fac-símiles, pelo que se justifica, neste ponto, relembrar os tipos de navegação identificados por alguns autores (Marchionini, 1995, p. 100 e ss.; Large et al., 2001, p. 181–182), a saber:

- navegação direta ou específica, ou seja, sistemática e direcionada para um determinado objetivo,
- navegação semi-direta e preditiva, que tem um alvo menos definido e é executada de forma menos sistemática,
- navegação indireta ou geral, a qual não tem um objetivo determinado.

Não há muitos estudos dedicados ao conhecimento da forma como os investigadores, sejam historiadores ou de outras áreas científicas, utilizam as coleções de jornais históricos. No entanto, os trabalhos de alguns autores permitem-nos compreender como os utilizadores de determinadas comunidades académicas o executam, podendo encontrar-se algumas diferenças entre utilizadores de diferentes comunidades.

Allen e Sieczkiewicz (2010) levaram a cabo um estudo com vista ao conhecimento das práticas de pesquisa e utilização de jornais históricos, independentemente do formato, ou seja, incluindo jornais em papel e jornais digitalizados. Este estudo teve como método de recolha de dados o recurso a entrevistas, pelo que é realçada a profundidade dos dados recolhidos e consequente caraterização das práticas dos entrevistados, neste caso todos os historiadores norte-americanos. Como resultado do estudo, os autores concluíram que os jornais históricos eram utilizados, sobre-

tudo, para a verificação de factos (nomes, datas e locais) e para a recolha de dados sobre temas mais amplos, como, por exemplo, eleições. Os entrevistados também relataram a utilização de jornais para preencher lacunas na investigação e corroborar a informação de outras fontes. Quando à utilização de bases de dados de jornais digitalizados, os entrevistados apresentaram uma tendência para a navegação em oposição à pesquisa, ou seja, para a procura semelhante à praticada em jornais em suporte papel através da análise integral dos jornais. Quando pesquisam, interpretando-se pesquisa como a construção e inserção de *queries* de pesquisa, estes investigadores utilizam preferencialmente tópicos, datas ou nomes de pessoa.

O comportamento informacional dos utilizadores da base de dados Welsh Newspapers Online (WNO)[1] foi estudado por Gooding (2016a) através da recolha de dados webométricos e posterior análise recorrendo a ferramentas analíticas da Google. A coleção WNO compreende 15 milhões de artigos de jornais digitalizados com datas que se situam entre 1804 e 1919. O autor concluiu que os utilizadores consultam primordialmente a página de título mais do que qualquer outra parte dos jornais e que, provavelmente, isto é um reflexo do modo como o significado formal da página de rosto foi reforçado pela interface de navegação, pois os utilizadores que acedem a jornais através da navegação, por exemplo, serão levados por omissão para a página de título. No entanto, neste mesmo estudo, o autor verifica que é significativo o facto de os utilizadores não dedicarem a mesma atenção às restantes páginas, aparentemente por confiarem na tecnologia em vez de procederem a uma navegação manual dos jornais. Gooding (2016a) vai mais longe e afirma que a dependência de ferramentas de filtragem automatizada é inevitável num recurso de grande escala e na pesquisa num largo volume de jornais. O facto de os utilizadores não navegarem através de edições sequenciais não sugere uma diminuição do interesse, até porque se está a assistir a um tipo de comportamento que combina pesquisa, navegação e leitura numa plataforma web.

Num estudo mais recente, Chardonnens et al. (2018) recorreram a ferramentas analíticas da web para analisar o conteúdo das pesquisas dos utilizadores na plataforma de jornais históricos em linha da Biblioteca Real da Bélgica. A BelgicaPress[2] disponibiliza o acesso a 112 jornais de 1814 a 1970, em francês, alemão e neerlandês, num total de cerca de 3,7 milhões de páginas. Os autores concluíram que é frequente a pesquisa por nomes de pessoa e por lugares. Esta afirmação vem ao encontro do identificado no estudo de Allen e Sieczkiewicz (2010), refletindo a necessidade de os utilizadores terem acesso a instrumentos de pesquisa adequados às suas práticas de investigação.

Bogaard et al. (2018), por seu lado, analisaram os *logs* de pesquisa combinados com registos de metadados descritivos do conteúdo da coleção de jornais históricos da Biblioteca Nacional da Holanda[3], utilizando estes metadados para criar subconjuntos nos *logs* correspondentes a diferentes partes da coleção. Os documentos da coleção analisada têm registos bibliográficos com os seguintes metadados: identificador de documento, data de publicação, tipo de artigo, título do jornal, local de publicação, fonte (a localização física do documento original) e zona de distribuição, a qual pode ter a

[1] Disponível em: https://newspapers.library.wales/home/

[2] Disponível em: https://www.belgicapress.be/

[3] Disponível em: https://www.delpher.nl/

identificação de "local", "nacional", de uma das antigas colónias holandesas ou, ainda, "desconhecido". Esta coleção inclui 1.500 títulos de jornais com datas que se situam entre 1618 e 1995, podendo recuperar-se jornais completos, páginas de jornais ou artigos individuais, considerando, neste caso, quatro tipos: artigos noticiosos, anúncios publicitários, anúncios (relativos à família, tais como anúncios de nascimento, casamento ou morte) e imagens. Os autores deste estudo concluíram que os utilizadores pesquisam por *queries* curtas, maioritariamente compostas por dois termos, recorrem pouco aos operadores booleanos e reordenam frequentemente os resultados por data.

Através do recente trabalho das autoras Late e Kumpulainen (2021) é possível perceber de que modo os utilizadores desenvolvem a pesquisa e interagem com as plataformas digitais de jornais históricos disponibilizadas pela Biblioteca Nacional da Finlândia. O objetivo do trabalho foi o de estudar qualitativamente as interações dos investigadores de história no âmbito dos jornais finlandeses digitalizados e a sua utilização como fontes primárias de investigação — ou seja, como fontes de dados de investigação que forneçam testemunhos em primeira mão ou evidencia diretamente relacionados com o tema da investigação histórica que desenvolvem. A Biblioteca Nacional da Finlândia possui jornais históricos digitalizados publicados na Finlândia entre 1771 e 1929, disponíveis tanto para cidadãos como para académicos. Trata-se de uma coleção com aproximadamente 7,4 milhões de páginas de jornais, nas línguas finlandesa, sueca e russa[4]. Uma das conclusões mais relevantes deste estudo reside no facto de ter sido identificada uma mudança na forma como os investigadores passaram a trabalhar a partir do uso de coleções digitais: a tradição de investigação baseada em trabalho individual está a transformar-se no sentido de processos mais colaborativos. Com o trabalho frequentemente desenvolvido por equipas de investigação multidisciplinares, surge a necessidade de um maior conhecimento de como a colaboração multidisciplinar se poderá desenvolver com base no recurso às tecnologias da informação para a investigação histórica. O estudo de Late e Kumpulainen (2021) destaca-se pelo detalhe da descrição das pesquisas desenvolvidas pelos entrevistados. As autoras verificaram que alguns tópicos não eram passíveis de pesquisa por palavra-chave e para procurá-los os investigadores recorreram à navegação pelos jornais, página a página, método que utilizam por não confiarem totalmente no reconhecimento ótico de carateres (OCR). Trata-se de um aspeto relevante e será novamente abordado, mais à frente no presente artigo.

Estas conclusões merecem uma reflexão, pois evidenciam a necessidade de desenho das bases de dados de jornais digitalizados com recurso à pesquisa assistida por sistemas de organização do conhecimento, no sentido de fornecer ao utilizador a terminologia mais adequada.

Potencialidades e limitações da digitalização dos jornais

As vantagens da digitalização de jornais históricos são irrefutáveis, particularmente por viabilizarem, a muitos utilizadores em simultâneo, o acesso a fontes que de outra forma estaria comprometido. Apesar de existirem bases de dados que re-

[4] Disponível em: https://digi.kansalliskirjasto.fi/

querem pagamento de uma subscrição para permitirem o acesso de investigadores e leitores em geral (Popik, 2004), existem cada vez mais projetos de digitalização de jornais históricos que permitem o acesso livre, como os exemplos apresentados no ponto anterior, ou, ainda, como o caso do projeto BC [British Columbia] Historical Newspapers[5] ou do Chronicling America[6]. Estes projetos constituem recursos que, como afirma Mussell (2012, p. 58) oferecem uma solução técnica para os problemas bibliográficos e metodológicos colocados pelas publicações impressas. A digitalização pode reunir séries fragmentadas e distribuídas por diversas bibliotecas e arquivos no seu espaço digital.

O aumento dos projetos de acesso livre ou comerciais contribui para o aperfeiçoamento tecnológico de reconhecimento ótico de carateres e expande as possibilidades de mais projetos poderem ser desenvolvidos a menores custos. Por exemplo, o OCR permite a pesquisa no texto de cada notícia digitalizada e de cada página. Para além do acesso ao texto completo e da pesquisa baseada em palavras-chave, as funções típicas incluem, ainda, a disponibilização de metadados, a navegação e a filtragem dos resultados. As interfaces mais avançadas oferecem, também, a funcionalidade de enriquecimento do conteúdo, como a correção pós-OCR, e ações de conectividade, como, por exemplo, as ligações a outros repositórios (Late e Kumpulainen, 2021).

Apesar de toda a evolução ocorrida neste âmbito, persistem, ainda, algumas críticas aos resultados da digitalização. Kettunen et al., (2014), Kettunen e Pääkkönen, (2016) e, ainda, Jarlbrink e Snickars (2017) identificaram, em algumas bases de dados, erros na digitalização por OCR que comprometiam a leitura das notícias e conduziam a textos que nunca tinham sido redigidos. Järvelin et al. (2016) apontam dois problemas significativos na transferência de recursos impressos para digitalizados. Enquanto a leitura por OCR pode atualmente alcançar mais de 99% de precisão no reconhecimento de caracteres de imagens de alta qualidade de documentos originais como os livros, no caso dos jornais históricos a precisão pode ser muito inferior. Os autores indicam, ainda, que a qualidade do OCR depende do ambiente e do estado de conservação dos documentos originais, ou seja, a qualidade da impressão e do papel. Os tipos de letra utilizados nos originais e a complexidade do *layout* também afetam a exatidão do resultado, pelo que, geralmente, quanto mais antigo for o jornal, menor será a taxa de exatidão. É, no entanto, possível encontrar projetos de aperfeiçoamento de deteção do *layout* recorrendo a componentes de OCR em acesso aberto, para manter um baixo custo do processo de digitalização (Liebl e Burghardt, 2020). O segundo problema identificado por Järvelin et al. (2016) diz respeito à mudança histórica nas línguas: os textos digitalizados são escritos na língua da época da impressão. Este aspeto apenas é ultrapassado quando está disponível uma correspondência terminológica, o que requer mais recursos e investimento na digitalização e disponibilização para consulta de uma coleção digitalizada. A correção dos erros de OCR também é reconhecidamente um processo moroso, por envolver um

[5] O BC [British Columbia] Historical Newspapers disponibiliza as versões digitalizadas de jornais dessa província canadiana com datas compreendidas entre 1865 e 1994. Está disponível em: https://open.library.ubc.ca/collections/bcnewspapers/

[6] O Chronicling America permite o acesso e pesquisa em jornais dos Estados Unidos relativos a um período temporal que se situa entre 1777 e 1963. Está disponível em: https://chroniclingamerica.loc.gov/

trabalho manual na obtenção de melhores resultados, como identificaram Strange et al. (2014) num estudo em que atingiram 98% de precisão com a introdução de correção manual pós-OCR.

Observando todos estes aspetos, e citando Mussell (2012, p. 31), pode concluir-se que:

> A forma como muitos destes recursos são construídos pressupõe que o que os utilizadores querem são artigos sobre algo, e este 'algo' é redutível à informação verbal na página. A utilização de transcrições textuais privilegia o que muda de artigo para artigo — a informação linguística — ignorando, ao mesmo tempo, as características formais que estabelecem os limites desta variação. São estas características formais, repetidas em cada fascículo, que situam os artigos dentro de fascículos, tiragens, publicações e cultura impressa de forma mais ampla[7].

Interfaces do utilizador, metadados descritivos e controlo de autoridades

O desenho das interfaces digitais de jornais históricos tem experimentado alguma evolução, após uma fase inicial em que as características essenciais se centravam na disponibilização de imagens e em rudimentares opções de pesquisa (Pfanzelter et al., 2021).

Ehrmann et al. (2019) desenvolveram uma análise de 24 interfaces do utilizador de jornais históricos digitalizados com origem australiana, europeia e norte-americana. Para levarem a cabo a análise, os autores deste estudo desenvolveram uma lista de características das interfaces do utilizador, distribuídas por categorias gerais e correspondentes propriedades.

Os autores identificaram quatro gerações de interfaces: a primeira centrada principalmente na disponibilização de conteúdos em linha, a segunda na interação avançada do utilizador com o conteúdo, a terceira no enriquecimento automático através do processamento de linguagem natural e a quarta na personalização dos serviços e maior transparência relativamente à composição do corpus e exploração visual, estando esta última geração ainda em desenvolvimento em projetos de investigação. No conjunto de interfaces analisadas, os autores verificaram que muitas das propriedades presentes na lista elaborada aparecem em menos de metade das interfaces. As características mais presentes incluem metadados dos jornais, modalidades de pesquisa, filtragem de resultados e visualização do conteúdo dos jornais (por exemplo, fac-símile ou vista geral dos fascículos disponíveis), enquanto as menos utilizadas englobam o enriquecimento semântico, a conectividade (por exemplo, ligação a outros repositórios), a informação sobre a digitalização e as APIs (Interfaces de Programação de Aplicações). Os resultados deste estudo demonstram, portanto, que ainda há espaço para mais desenvolvimento em torno das plataformas que disponibilizam jornais históricos digitalizados.

Mais recentemente, Pfanzelter et al. (2021), partindo da análise de três estudos anteriores, concluíram que, até agora, a utilização das interfaces dos jornais é complexa e, de algum modo, restritiva, devido a fatores como a qualidade da leitura por OCR, a imperfeição das imagens e a presença de metadados incorretos, bem como devido

[7] Tradução da autora

à ausência de funcionalidades para análise e visualização. As ferramentas existentes que não estão integradas nas interfaces carecem frequentemente de transparência, de facilidade de utilização e da possibilidade de processar grandes corpora de dados.

Pela sua exaustividade, o conjunto de propriedades identificado por Ehrman et al. (2019), no estudo acima referido, pode servir de base para a especificação de requisitos subjacente ao desenvolvimento de bases de dados de acesso e pesquisa em jornais históricos digitalizados. Algumas das propriedades exigem um maior esclarecimento, sobretudo quanto à identificação dos metadados descritivos e de controlo de autoridades, pelo que, no presente artigo, se desenvolveu a sua adaptação, apresentada no Quadro 1. Esta adaptação considerou a revisão do vocabulário usado por aqueles autores à luz dos princípios utilizados na descrição bibliográfica normalizada e na organização do conhecimento, tradicionalmente realizadas no contexto biblioteconómico e, atualmente, aplicadas em diversos contextos. Ou seja, procedeu-se à aplicação da terminologia veiculada na norma internacional de descrição bibliográfica editada pela IFLA[8], bem como na terminologia presente nas normas editadas pela ISO relativamente ao desenvolvimento de thesauri e interoperabilidade com outros vocabulários (IFLA, 2011; ISO, 2011; ISO, 2013). Estas normas divulgam os princípios gerais da determinação dos pontos de acesso por autor, título e assunto, devendo ser complementadas, por exemplo, com a consulta a dados de autoridade de nome, disponíveis no VIAF[9].

Para desenvolver o quadro 1, também se procedeu à fusão entre algumas categorias e à revisão de algumas propriedades que se podem considerar de relevância secundária, em função das práticas de utilização dos jornais históricos já abordadas no presente artigo, bem como de propriedades identificadas nas observações de Pfanzelter et al. (2021).

O conjunto de propriedades a considerar no desenvolvimento de uma coleção digitalizada distribui-se, então, por seis categorias, a saber:

- Informação sobre a interface e documentação sobre a digitalização e conectividade dos conteúdos — propriedades que identificam e caraterizam a interface e conferem maior transparência quanto às funcionalidades técnicas;
- Informação sobre a coleção e metadados dos jornais — propriedades que ilustram a caraterização geral da coleção, bem como dos jornais e seus conteúdos;
- Navegação, opções de pesquisa e filtragem dos resultados — propriedades essenciais à navegação entre os jornais e suas páginas e funcionalidades de pesquisa que permitem recuperar os artigos e os jornais pelos mais diversos pontos de acesso, bem como reformular a estratégia de pesquisa;
- Visualização e ordenação dos resultados da pesquisa — propriedades que indicam a distribuição dos resultados da pesquisa e a sua (re)ordenação;
- Visualização do conteúdo dos jornais — propriedades que apontam para as variadas opções de visualização do conteúdo dos jornais e das notícias fruto da navegação ou dos resultados da pesquisa efetuada;

[8] Acrónimo de International Federation of Library Associations and Institutions

[9] Acrónimo de Virtual International Authority File, disponível em: http://viaf.org/

— Área reservada e interação dos utilizadores — propriedades que auxiliam o utilizador no uso, conservação e manuseio dos resultados da sua pesquisa.

Na tabela seguinte, visualiza-se, deste modo, um conjunto de categorias diretamente relacionadas com a especificidade do tipo de material em causa e tipo de conteúdo, ou seja, os jornais históricos e os seus artigos, e categorias mais relacionadas com as funcionalidades técnicas das estruturas das bases de dados.

Quadro 1. Critérios e propriedades a implementar no desenvolvimento ou avaliação de coleções de jornais históricos digitalizados

Categorias	Propriedades
Informação sobre a interface e documentação sobre a digitalização e conectividade dos conteúdos	• URL • Área alvo • Finalidade e alcance • Criador da interface • Data da criação • Línguas da interface • Modelo de acesso • Fornecedor da interface • Documentação sobre a disposição a nível do artigo • Documentação sobre enviesamentos e falhas • Pontuação de relevância do resultado da pesquisa • Data de digitalização ao nível dos títulos • Resolução da digitalização (em dpi) • Informação sobre as ferramentas de OCR utilizadas • Aviso de direitos de autor • Documentação sobre os métodos da digitalização • Identificadores de terceiros (por exemplo, VIAF) • Ligações a outros repositórios • Tecnologias da web semântica • Ligação ao código fonte da interface • APIs IIIF
Informação sobre a coleção e metadados dos jornais	• Número total de títulos de jornais • Número total de fascículos • Número total de páginas • Número total de artigos • Indicação da edição digitalizada original • Novos títulos continuamente adicionados • Línguas dos jornais • Títulos próprios, títulos alternativos, títulos relacionados, títulos anteriores, títulos subsequentes dos jornais • Local de publicação • Cobertura geográfica • Editora • Intervalo de datas • Periodicidade • ISSN • Descrição do jornal (contexto) • Idioma • Vista de calendário dos fascículos • Indicação do detentor do arquivo/biblioteca • Tema do jornal (representação através de descritores dos vocabulários controlados) • Tema do artigo (representação através de descritores dos vocabulários controlados) • Entidades: nome de pessoa, nome de coletividade, nome de família • Tipo de conteúdo (anúncio, artigo, ilustração)

Navegação, opções de pesquisa e filtragem dos resultados	• Navegação por data • Navegação por título • Navegação por local de publicação • Navegação por tema do jornal (ligação aos metadados) • Navegação por tema do artigo (ligação aos metadados) • Pesquisa básica por palavra-chave • Preenchimento automático • Operadores booleanos (AND, OR, NOT) • Pesquisa de frases • Pesquisa por truncatura • Pesquisa por operadores de proximidade • Limite de intervalo de datas • Limite por língua • Limite por título de jornal • Limite por local de publicação • Limite por tema do jornal • Limite por tema do artigo • Pesquisa por entidades: nome de pessoa, nome de coletividade, nome de família • Pesquisa por nome de lugar • Limite por segmentos / zonas dos jornais • Limite por tipo de conteúdo (anúncio, artigo, ilustração) • Limite por comprimento do artigo • Limite por detentor do arquivo / biblioteca • Limite por licença / acessibilidade • Limite por frequência de publicação • Limite por data de disponibilização online • Sugestão de pesquisa
Visualização e ordenação dos resultados da pesquisa	• Visualização da distribuição ao longo do tempo (cronológica e em períodos temporais) • Visualização da distribuição por local de publicação • Visualização da distribuição por tema do jornal • Visualização da distribuição por tema do artigo • Destaque de palavras-chave em fac-símiles • Destaque de palavras-chave no texto lido por OCR • Ordenação por relevância • Ordenação por data • Ordenação por título de jornal • Ordenação por título de artigo • Ordenação por tipo de conteúdo (anúncio, artigo, ilustração) • Ordenação por data de publicação online • Ordenação por língua • Ordenação por popularidade (número de visualizações)
Visualização do conteúdo dos jornais	• Disponibilização do fac-símile • Visualização de texto lido por OCR • Visualização da largura/altura total real da página • Disponibilização da miniatura interactiva da página • Visão geral dos fascículos disponíveis • Pesquisa na página visualizada • Opção de continuar para a página seguinte • Opção de continuar para o resultado seguinte

Área reservada e interação dos utilizadores	• Guardar artigos numa pasta pessoal • Guardar as estratégias de pesquisa numa pasta pessoal • Manter um registo dos materiais visualizados • Ligação permamente do jornal • Exportação da referência bibliográfica • Opção para corrigir OCR • Possibilidade de adicionar/editar metadados • Ferramenta de captura de ecrã • Opções de descarga (formatos de ficheiro) • Possibilidade de organizar artigos em coleções pessoais • Correção pós-OCR • Reutilização de texto • Recomendações

Fonte: elaboração da autora (adaptado de Ehrmann et al., 2019, e de Pfanzelter et al., 2021)

Da observação da tabela, pode-se, então, compreender que é vasto o conjunto de propriedades, sendo que, principalmente em algumas categorias, como: Informação sobre a coleção e metadados dos jornais, Visualização do conteúdo dos jornais e Área reservada e interação dos utilizadores, possa haver no futuro a introdução de mais propriedades no seguimento da evolução tecnológica, quer quanto aos procedimentos de digitalização, quer relativamente às possibilidades de hiperligação a fontes externas de dados.

O quadro apresentado poderá, por outro lado, servir de lista de verificação para uma avaliação de plataformas e bases de dados de disponibilização de jornais históricos digitalizados, individuais ou coleções. A atualização das categorias e correspondentes propriedades deve realizar-se periodicamente, dado que a evolução tecnológica e o aparecimento de mais aplicações de análise de texto que, no atual momento, se encontram ainda em desenvolvimento, podem contribuir para alterações significativas, sobretudo, no que respeita à pesquisa e análise de conteúdo.

Conclusão

Neste artigo apresentou-se um conjunto de estudos sobre o comportamento informacional dos utilizadores de determinadas comunidades académicas, no sentido de se evidenciar as preferências desses utilizadores no uso dos jornais históricos digitalizados e nas práticas de pesquisa. Após uma breve análise das potencialidades e limitações da digitalização dos jornais, abordou-se, ainda, as interfaces do utilizador e identificou-se, a partir de estudos anteriores, um conjunto de caraterísticas, distribuídas por categorias e propriedades, essenciais ao desenvolvimento e avaliação de plataformas e bases de dados de disponibilização de jornais históricos digitalizados.

O desenvolvimento técnico das bases de dados e a apresentação de um completo conjunto de funcionalidades dependerá, sempre, do investimento possível de alcançar e dos recursos que estes processos de digitalização de materiais impressos exigem. Sobretudo, se o que está em causa é o acesso livre por parte dos utilizadores, ou seja, sem necessidade de, quer os investigadores, quer as suas instituições de origem, subscreverem o acesso às bases de dados.

Dos estudos existentes sobre o comportamento informacional dos utilizadores, bem como dos trabalhos publicados mais recentemente sobre o desenvolvimento e análise das interfaces do utilizador, sobressai a necessidade de um maior investi-

mento na disponibilização de meios automatizados que assistam os investigadores na pesquisa e acesso aos jornais históricos digitalizados e que lhes permitam, pela facilidade no uso e pela transparência dos resultados decorrentes das aplicações tecnológicas, desenvolver novos métodos de investigação e encontrar novas interpretações para os seus objetos de estudo. Deste modo, será relevante acompanhar a evolução dos estudos experimentais em curso sobre, por exemplo, o processamento de linguagem natural aplicado aos jornais históricos, o reconhecimento automático de entidades através de ligação a ficheiros de autoridade ou as aplicações para análise de sentimento.

A finalizar, sublinha-se a necessidade de se prosseguir com o estudo do comportamento informacional de utilizadores de jornais históricos digitalizados, sobretudo em diferentes comunidades, académicas e não académicas, para se possa compreender as alterações aos padrões de investigação e a consolidação das práticas de pesquisa. É, pois, o conhecimento aprofundado das necessidades informacionais dos utilizadores que permitirá fundamentar a exploração de mais funcionalidades técnicas dos recursos em análise neste artigo e contribuir para a definição das melhores práticas de agregação e disponibilização de jornais históricos digitalizados.

Referências Bibliográficas

Allen, R. B., & Sieczkiewicz, R. (2010). How historians use historical newspapers. *Proceedings of the American Society for Information Science and Technology*, *47*(1), 1–4. https://doi.org/10.1002/meet.14504701131

Bogaard, T., Hollink, L., Wielemaker, J., Van Ossenbruggen, J., & Hardman, L. (2019). Metadata categorization for identifying search patterns in a digital library. *Journal of Documentation*, *75*(2), 270–286.
https://doi.org/10.1108/JD-06-2018-0087

Brake, L. (2012). Half Full and Half Empty. *Journal of Victorian Culture*, *17*(2), 222–229. http://dx.doi.org/10.1080/13555502.2012.683149

Chardonnens, A., Rizza, E., Coeckelbergs, M., & van Hooland, S. (2018). Mining user queries with information extraction methods and linked data. *Journal of Documentation*, *74*(5), 936–950.
https://doi.org/10.1108/JD-09-2017-0133

Ehrmann, M., Bunout, E., & Düring, M. (2019). Historical newspaper user interfaces: a review. *IFLA WLIC 2019 - Athens, Greece - Libraries: Dialogue for Change*. http://library.ifla.org/id/eprint/2578/1/085-ehrmann-en.pdf

Gooding, P. (2016a). Exploring the information behaviour of users of Welsh Newspapers Online through web log analysis. *Journal of Documentation*, *72*(2), 232–246. https://doi.org/10.1108/JD-10-2014-0149

Gooding, P. (2016b). *Historic Newspapers in the Digital Age: Search All about it*! Routledge.

IFLA (2011). *ISBD International Standard Bibliographic Description — Consolidated edition*. De Gruyter Saur.
https://repository.ifla.org/handle/123456789/786

International Organization for Standardization. (2011). *Information and documentation — Thesauri and interoperability with other vocabularies — Part 1: Thesauri for information retrieval*

(ISO Standard No. 25964-1:2011). https://www.iso.org/standard/53657.html
International Organization for Standardization. (2013). *Information and documentation — Thesauri and interoperability with other vocabularies — Part 2: Interoperability with other vocabularies* (ISO Standard No. 25964-2:2013). https://www.iso.org/standard/53658.html
Hansen, K. A., & Paul, N. (2015). *News Archive Chaos: A Case Study* [Paper present]. IFLA News Media and Audiovisual and Multimedia Sections' Conference.
https://www.ifla.org/files/assets/newspapers/Sweden_2015/6_-_hansen_and_paul_ifla_2015_news_archive_chaos.pdf
Hitchcock, T. (2013). Confronting the Digital Or How Academic History Writing Lost the Plot. *Cultural and Social History*, *10*(1), 9-23. https://doi.org/10.2752/147800413X13515292098070
Jarlbrink J., & Snickars, P. (2017). Cultural Heritage as Digital Noise: Nineteenth Century Newspapers in the Digital Archive. *Journal of Documentation*, *73*(6), 1228-43. https://doi.org/10.1108/JD-09-2016-0106
Järvelin, A., Keskustalo, H., Sormunen, E., Saastamoinen, M., & Kettunen, K. (2016). Information retrieval from historical newspaper collections in highly inflectional languages: A query expansion approach. *Journal of the Association for Information Science & Technology*, *67*(12), 2928-2946. https://doi.org/10.1002/asi.23379
Kettunen, K., Honkela, T., Lindén, K., Kauppinen, P., Pääkkönen, T., & Kervinen, J. (2014). *Analyzing and improving the quality of a historical news collection using language technology and statistical machine learning methods*. IFLA 80th World Library and Information Congress Proceedings. http://hdl.handle.net/10138/136269
Kettunen, K., & Pääkkönen, T. (2016). Measuring lexical quality of a historical Finnish newspaper collection — analysis of garbled OCR data with basic language technology tools and means. *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (pp. 956-961). https://www.aclweb.org/anthology/L16-1152.pdf
Late, E., & Kumpulainen, S. (2021). Interacting with digitised historical newspapers: understanding the use of digital surrogates as primary sources. *Journal of Documentation*, vol. ahead-of-print. No. ahead-of-print. https://doi.org/10.1108/JD-04-2021-0078
Large, A., Tedd, L., & Hartley, R.J. (2001). *Information seeking in the online age: principles and practice*. K.G. Saur.
Liebl, B., & Burghardt, M. (2020). *From Historical Newspapers to Machine-Readable Data: The Origami OCR Pipeline* [Conference presentation]. CHR 2020: Workshop on Computational Humanities Research. http://ceur-ws.org/Vol-2723/long20.pdf
Marchionini, G. (1995). *Information Seeking in Electronic Environments*. Cambridge University Press.
Mussell, J. (2012). *The Nineteenth-Century Press in the Digital Age*. Palgrave Macmillan.
Pfanzelter, E., Oberbichler, S., Marjanen, J., Langlais, P., & Hechl, S. (2021). Digital interfaces of historical newspapers: opportunities, restrictions and recommendations. *Journal of Data Mining & Digital Humanities*, In press, HistoInformatics. hal-02480654v5
https://doi.org/10.46298/jdmdh.6121
Popik, B. (2004). Digital Historic Newspapers: A Review of Powerful New Research Tools. *Journal of English Linguistics*, *32*(2), 114-123. https://doi.org/10.1177/0075424204265818
Strange, C., Wodak, J., & Wood, I. (2014). Mining for the Meanings of a Murder: The Impact of OCR Quality on the Use of Digitized Historical Newspapers. *Digital Humanitites Quarterly*, *8*(1).
http://www.digitalhumanities.org/dhq/vol/8/1/000168/000168.html

Nota biográfica

Olívia Pestana é Professora Auxiliar do Departamento de Ciências da Comunicação e da Informação da Faculdade de Letras da Universidade do Porto. É investigadora do CITCEM — Centro de Investigação Transdisciplinar Cultura, Espaço e Memória.

ORCID ID: 0000-0002-5485-3143

Ciência ID: A319-7793-ACB0

Scopus ID: 56926787100

Morada institucional: Faculdade de Letras da Universidade do Porto, Via Panorâmica, s/n - 4150-564 Porto, Portugal

How to cite:

Pestana, O. (2021). O acesso aos jornais históricos: Considerações sobre o desenvolvimento de coleções digitalizadas. *Revista Media & Jornalismo, 21*(39), 162–174. https://doi.org/10.14195/2183-5462_39_8

Submetido | Received: 2021.03.17

Aceite | Accepted: 2021.05.20