A utilização de sistemas de organização do conhecimento em fontes para o estudo da história do jornalismo

The use of knowledge organization systems in sources for the study of the history of journalism

Olívia Pestana

Universidade do Porto / Faculdade de Letras /CITCEM
opestana@letras.up.pt
ORCID ID: 0000-0002-5485-3143

Resumo: A criação e desenvolvimento de bases de dados de jornais históricos é essencial para o estudo da história do jornalismo. São vários os projetos internacionais que se têm dedicado a este propósito, havendo um especial enfoque na digitalização, disponibilização e acesso, colocando, não raras vezes, em segundo plano o devido controlo de autoridade por assuntos. É marcante, neste âmbito, a quase ausência da indexação por assuntos normalizada, não permitindo, por este motivo, o desenvolvimento de processos de interoperabilidade com outras bases de dados relativamente à pesquisa por assunto. O controlo do vocabulário existente nos sistemas de organização do conhecimento visa reduzir a ambiguidade da linguagem natural e contribuir para a recuperação de informação relevante diante de uma determinada necessidade de informação. Este trabalho começa por dar a conhecer e analisar criticamente alguns projetos internacionais relevantes de coleções de jornais históricos digitalizados, descrevendo o percurso subjacente ao desenvolvimento técnico das respetivas bases de dados e apresentando as diversas formas de recuperação da informação, salientando as possibilidades de pesquisa por assunto. São apresentadas diversas possibilidades de utilização dos sistemas de organização do conhecimento na indexação de jornais históricos, sendo, também, exemplificada a sua aplicabilidade em coleções de jornais portugueses.

Palavras-chave: indexação por assuntos; sistemas de organização do conhecimento; fontes de informação; jornais históricos; história do jornalismo.

Abstract: *The creation and development of databases of historical newspapers are essential for studying the history of journalism. Several international projects have been dedicated to this purpose, focusing on digitization, availability and access. Often, the control of authority by the subject is not considered. It is striking, in this context, the absence of standardized subject indexing, which does not allow, for this reason, the development of interoperability processes with other databases concerning the search by subject. The control of the vocabulary existing in knowledge organization systems aims to reduce the ambiguity of natural language and contribute to the retrieval of relevant information in the face of a determined information need. This work begins by presenting and critically analyzing some relevant international projects from collections of digitized historical newspapers, explaining the path underlying the technical development of the databases and the various forms of information retrieval. Several possibilities of using knowledge organization systems in the indexing of historical newspapers are presented, showing their applicability also in Portuguese newspapers' collections.*

Keywords: *subject indexing; knowledge organization systems; information sources; historical newspapers; history of journalism.*

Introdução

A relevância do acesso aos jornais históricos para o estudo da história do jornalismo, bem como a afirmação destes jornais como fonte de informação para a investigação em áreas disciplinares como a história, as ciências da comunicação e da informação, ou outras áreas do âmbito das ciências sociais e humanas, promoveu o aparecimento de projetos de digitalização de jornais históricos integral ou parcelarmente. São muitos os projetos existentes, resultantes, em grande parte, de trabalhos individuais de formação pós-graduada ou integrados em centros de investigação, cujo propósito não é especificamente a digitalização dos jornais, mas sim o estudo de outros fenómenos. Este é, provavelmente, o principal motivo para a ausência da utilização de formatos normalizados de metadados descritivos e administrativos, que viabilizem a reutilização, a interoperabilidade, bem como o acesso a longo prazo aos jornais. Com efeito, são vários os projetos que, pouco tempo após a disponibilização dos dados recolhidos, não permitem o acesso aos resultados, constituindo, portanto, uma perda

de recursos e de acesso a dados que poderiam ser reutilizados e que constituem um património valioso e único.

Não obstante, encontram-se iniciativas de trabalho colaborativo de âmbito internacional, com planos de digitalização estruturada e normalizada, que têm contribuído para que o acesso aos jornais históricos se realize de forma permanente e com possibilidades de recuperação da informação recorrendo às mais recentes técnicas de pesquisa alicerçadas em resultados decorrentes da investigação em inteligência artificial (Popik, 2004, pp. 114-123). A par destas iniciativas, existem, também, bases de dados de cariz comercial que têm vindo a disponibilizar o acesso a jornais correntes e históricos, exigindo por parte dos utilizadores o pagamento dos acessos e do descarregamento dos jornais ou das páginas dos jornais (Mouhot, 2010, pp. 131-134).

Mas o aparecimento das bases de dados trouxe consigo uma transformação na observação dos jornais e das notícias que pode ter relevante impacto na forma de abordar a história do jornalismo. Hansen e Paul (2015, pp. 7-8) recuperam a análise de Gabriele (2014, p. 6) evidenciando como a transfiguração dos jornais, ao passar do papel ao microfilme e, depois, às bases de dados, retirou a leitura linear e cronológica dos jornais, largamente compreendida pelos leitores, viabilizando, no entanto, a identificação de pequenos jornais locais como fontes reconhecidas e, até então, minimizadas pelos utilizadores em geral. Mas, os autores vão mais longe ao considerar que as bases de dados, apesar desta vantagem, destroem em absoluto a ordem histórica da narrativa do jornal e facilitam a criação de enormes volumes de dados relativos a notícias de grandes regiões geográficas e períodos de tempo. Esta aspecto contribui para uma inevitável alteração da forma de pesquisa e análise dos jornais e das notícias, que viabilizará o caminho até novas formas de interpretar as questões de investigação. Conforme identificado por Birkner, Koenen & Schwarzenegger (2018, p. 1125), os historiadores dos *media* apenas começaram a descobrir e a explorar o potencial dos repositórios, fontes e métodos digitais para seu trabalho e a compreender as competências necessárias para tal.

Os requisitos dos processos de digitalização de jornais são conhecidos. Ao contrário dos atuais jornais que já são criados de forma digital, a transferência de suporte dos jornais da era da tipografia requer técnicas avanças que permitam o reconhecimento de textos, como, por exemplo, através do uso de software de reconhecimento ótico de caracteres (OCR) ou de reconhecimento inteligente de caracteres (ICR). Em bases de dados comerciais, é, ainda, aplicada a segmentação da leitura das páginas na digitalização e, pela utilização deste tipo de tecnologia, é possibilitada a leitura dos artigos e a pesquisa no seu conteúdo. A diferença no resultado da pesquisa é a de que o resultado a nível do artigo inclui o título do artigo, enquanto o resultado a nível da página inclui a data da edição e a página. Em qualquer das formas, a pesquisa recupera os textos que contêm a(s) palavra(s) ou frase(s) pesquisadas (Mouhot, 2010, pp. 131-134).

A recuperação de textos por assunto nas bases de dados de jornais históricos tem vindo a ser realizada recorrendo maioritariamente à utilização do reconhecimento da linguagem

natural utilizada na pesquisa e ao estabelecimento de correspondências com a lista de índices criados na leitura dos carateres. São escassos os exemplos de controlo de autoridade dos assuntos, ou seja, da disponibilização da pesquisa por termos incluídos em sistemas de organização do conhecimento. A linguagem natural é composta por um grande número de termos, sendo complexa, variada, caraterizada por casos de sinonímia ou quase sinonímia e, ainda, por situações de polissemia. Torna-se, portanto, necessário utilizar linguagens controladas para representar os conceitos, eliminando qualquer possibilidade de ambiguidade. Esta linguagem controlada constitui uma linguagem de representação dos recursos de informação, desenhada para facilitar a pesquisa. A conquista desta linguagem é a de estabelecer um acordo entre o sistema e o utilizador, pois serve para indexar o recurso de informação e também para indexar a pesquisa. No entanto, a modalidade de pesquisa por linguagem natural estará sempre presente e, em bases de dados com funcionalidades mais avançadas, poderá haver a correspondência automática entre os termos da linguagem natural aplicada pelo utilizador e os termos do vocabulário controlado usado na indexação dos recursos (Tartaglia, 2004, pp. 365-377).

É neste contexto que se torna relevante aprofundar a forma de tratamento por assunto das bases de dados e projetos de jornais históricos e a aplicabilidade dos sistemas de organização do conhecimento neste âmbito. Desenvolvemos, neste trabalho, o propósito de explorar algumas das bases e projetos de reconhecimento internacional, bem como analisar as possibilidades de utilização dos sistemas de organização do conhecimento, com vista a uma melhor recuperação da informação. O ponto de partida para a realização deste trabalho assenta na revisão de literatura concretizada a partir dos resultados da pesquisa sobre coleções de jornais históricos realizada em bases de dados de literatura científica, a saber: Communication Abstracts, Library and Information Science Source e, ainda Library and Information Science and Technology Abstracts. Para além destas bases, foram, ainda, consultadas obras monográficas de referência na área da indexação por assuntos e dos sistemas de organização do conhecimento direcionadas para o tratamento técnico de publicações periódicas.

Bases de dados e projetos internacionais de digitalização de jornais históricos

Neste ponto daremos a conhecer alguns projetos internacionais relevantes de coleções de jornais históricos digitalizados, apresentando o percurso subjacente ao desenvolvimento técnico das respetivas bases de dados e as diversas formas de recuperação da informação.

As bases de dados e os projetos apresentados foram selecionados com base no volume de jornais e de páginas de jornais digitalizadas, bem como pelas possibilidades de pesquisa oferecidas. Além destes aspetos, destacam-se por serem de acesso livre, bem como

permitirem o descarregamento das páginas dos jornais para reutilização livre não comercial, desde que estejam no domínio público.

Europeana Newspapers

A Europeana Newspapers[1] constitui uma das coleções disponibilizadas na plataforma Europeana e resulta de um projeto co-financiado pela União Europeia. Permite o acesso a 58 milhões de objetos digitais, incluindo livros, música, obras de arte, jornais e a outros recursos, como coleções temáticas, exposições, galerias e blogs. A Europeana Newspapers dá acesso a títulos, artigos, anúncios e artigos de opinião de jornais europeus de 23 países, com datas que se situam entre 1618 e 1996, perfazendo quase 5 milhões de páginas de jornais digitalizadas.

O projeto que deu origem a esta coleção de jornais digitalizados decorreu entre 2012 e 2015 e juntou 18 entidades participantes, com coordenação da Staatsbibliothek zu Berlin (Europeana, 2015).

Quanto às modalidades de recuperação da informação, é possível pesquisar por assunto no texto completo das páginas, sendo através da linguagem natural expressa na equação de pesquisa, ou seja, não são apresentadas possibilidades de utilização de vocabulário controlado para a pesquisa. No entanto, são apresentadas sugestões, de acordo a correspondência da palavra relativamente aos termos existentes (ver figuras 1 e 2).

Num projeto da dimensão da Europeana, será de elevada complexidade a disponibilização de termos de pesquisa em concordância com vocabulários controlados, dada a diferente origem dos jornais digitalizados e correspondentes metadados.

Relativamente aos metadados, é possível pesquisar por país de origem ou por título no menu de pesquisa disponível na primeira página, sendo, também, possível limitar a pesquisa efetuada a datas, língua ou, ainda, à instituição que fornece o acesso. Os esclarecimentos aos utilizadores relativamente às diversas modalidades de pesquisa são realizados numa página de esclarecimento geral sobre a Europeana, pelo que é exigido ao utilizador uma prática de familiarização com o sistema até conseguir obter resultados adequados à sua necessidade de informação. Por outro lado, são escassos, ou praticamente inexistentes, os estudos de caso relativos a recuperação da informação nesta plataforma. Tais estudos permitiriam conhecer as limitações concretas das possibilidades de pesquisa. No entanto, é de destacar que as modalidades apresentadas foram previamente testadas com grupos de utilizadores e recolhidas as propostas de melhoria a implementar (Willems & Atanassova, 2015, pp. 51-56).

1 Disponível em: https://classic.europeana.eu/portal/en/collections/newspapers

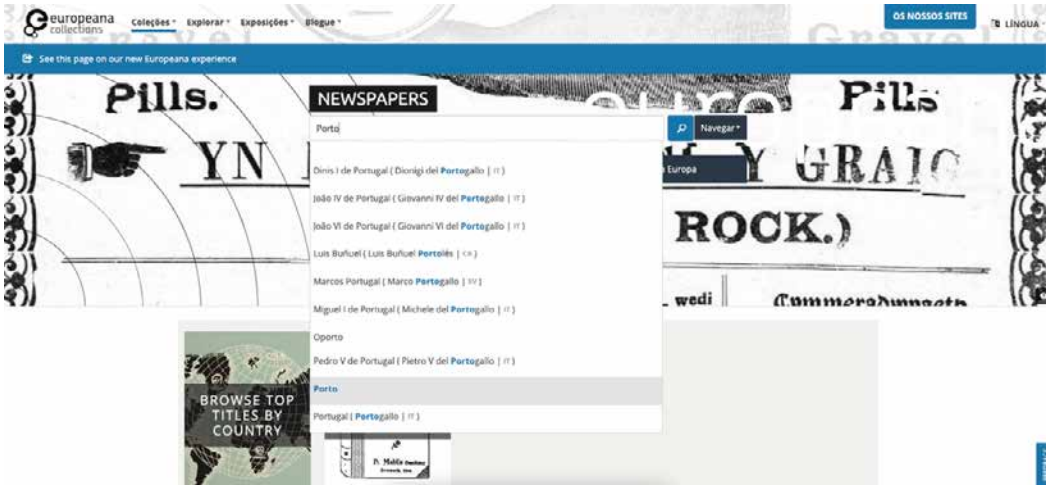

Figura 1
Exemplo 1 do menu de pesquisa por assunto.
Fonte: Europeana Newspapers.

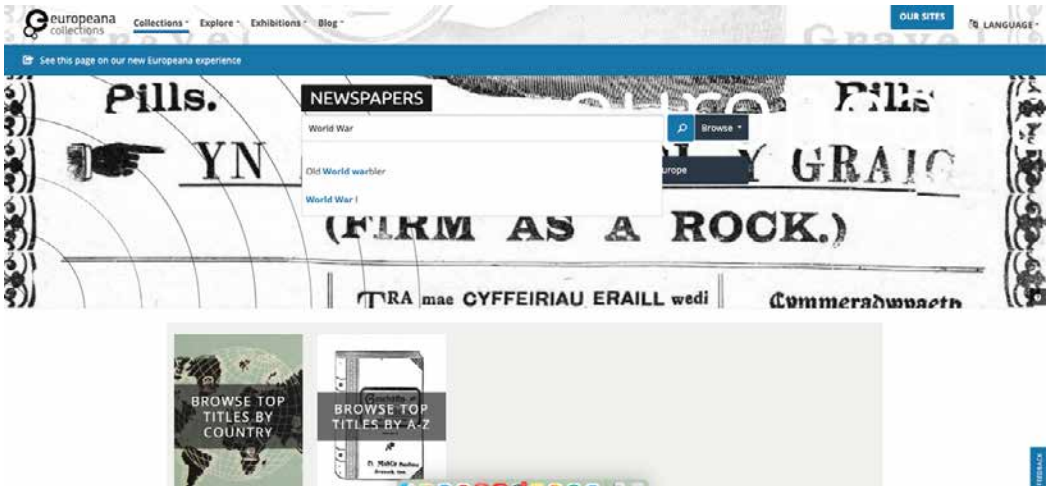

Figura 2
Exemplo 2 do menu de pesquisa por assunto.
Fonte: Europeana Newspapers.

Chronicling America

Chronicling America é um website que fornece acesso a jornais históricos e páginas selecionadas de jornais digitalizados e é produzido pelo National Digital Newspaper Program (NDNP)[2]. O NDNP resulta de uma parceria entre o National Endowment for the Humanities

2 Disponível em: https://chroniclingamerica.loc.gov/

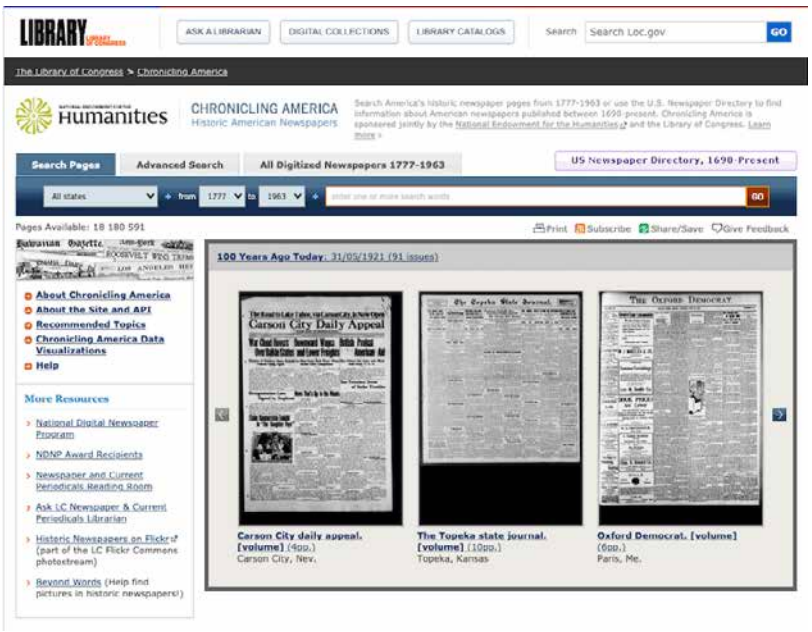

Figura 3
Página principal do website Chronicling America.
Fonte: Biblioteca do Congresso — Chronicling America.

(NEH) e a Biblioteca do Congresso, constituindo um esforço de longo prazo para fornecer acesso permanente a uma fonte digital nacional de informação bibliográfica de jornais correntes e de jornais históricos, selecionados e digitalizados por instituições financiadas pelo NEH de todos os estados e territórios dos EUA. Este programa baseia-se no legado do United States Newspaper Program, projeto que decorreu entre 1982 e 2011, patrocinado pelo NEH com suporte técnico da Biblioteca do Congresso, a qual organizou o inventário, a catalogação e a preservação seletiva em microfilme de um corpus constituído por materiais de jornais em risco de conservação.

O website Chronicling America apresenta, na sua página inicial, informação detalhada a respeito do projeto, bem como hiperligações para as diferentes modalidades de pesquisa inicial ou avançada (ver figura 3).

A pesquisa na base de dados realiza-se através de um alargado conjunto de possibilidades, como se pode visualizar na figura 4 ilustrativa do menu de pesquisa avançada. É possível pesquisar por palavras, limitando a um período temporal que se situa entre 1777 e 1963, selecionando um determinado estado norte-americano, ou, ainda, limitando a um dado título de jornal.

Estas possibilidades de pesquisa são viáveis num projeto deste cariz, dado que os metadados são preparados de acordo com os mesmos modelos desde a origem, ou seja, as Bibliotecas e demais entidades cooperantes seguem o modelo de preparação e exportação de dados em conformidade com a formatação exigida[3].

3 Ver Technical Guidelines for Applicants, disponível em: https://www.loc.gov/ndnp/guidelines/NDNP_202123TechNotes.pdf. Consultado em 30 abril de 2021.

Para um melhor sucesso na pesquisa por assunto, esta base de dados dá acesso aos tópicos e assuntos presentes nos artigos digitalizados. A lista de tópicos amplamente cobertos pela imprensa americana da época é apresentada noutra página web da Biblioteca do Congresso[4] hiperligada com a Chronicling America, sendo, ainda, possível observar os tópicos disponíveis em cada recorte temporal com diversas formas de visualização (ver figuras 5 e 6).

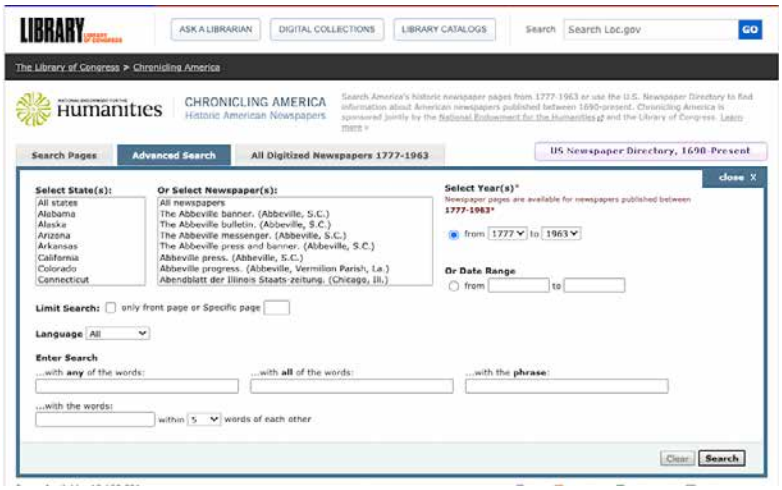

Figura 4
Imagem do menu de pesquisa avançada. Fonte: Biblioteca do Congresso — Chronicling America.

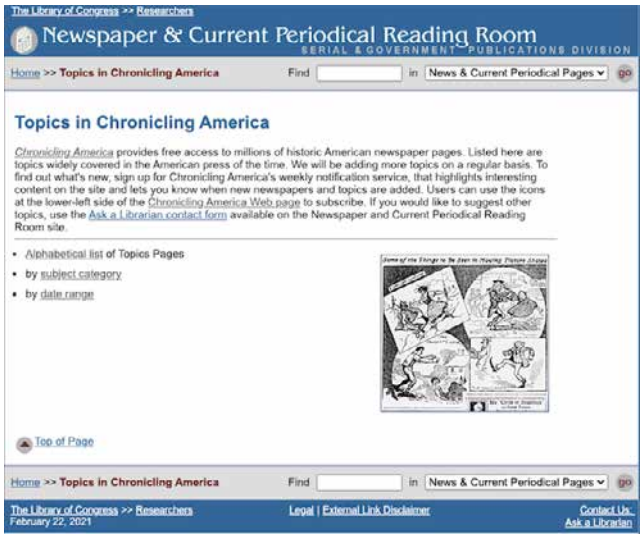

Figura 5
Imagem da página relativa à lista de tópicos. Fonte: Biblioteca do Congresso.

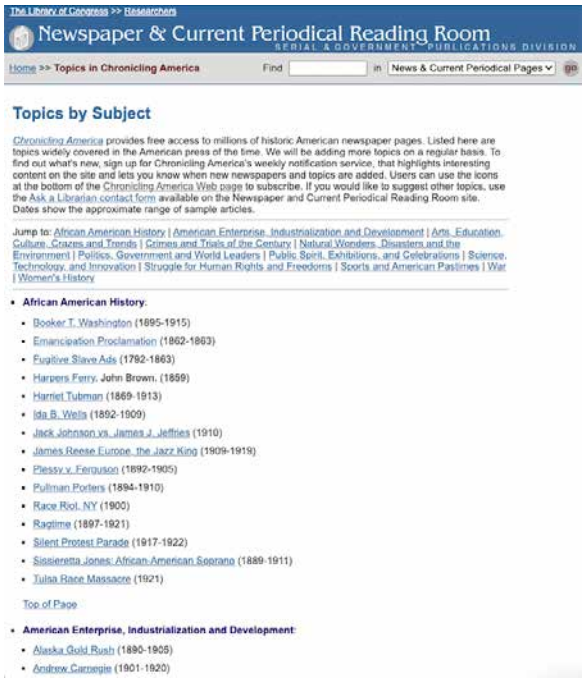

Figura 6
Imagem da lista de tópicos apresentada por categorias de assuntos. Fonte: Biblioteca do Congresso.

4 Ver: https://www.loc.gov/rr/news/topics/index.html

Conforme se pode constatar, este projeto revela uma substancial preocupação na abordagem ao acesso por assunto, constituindo, portanto, uma referência neste contexto. É de sublinhar que a Biblioteca do Congresso tem uma aprofundada experiência no desenvolvimento de linguagens controladas para o estabelecimento de pontos de acesso por assuntos. A primeira edição da lista de pontos de acesso atualmente intitulada Library of Congress Subject Headings surgiu em 1909, tendo presentemente ativa a sua 43ª edição[5]. Esta linguagem é utilizada em vários países integralmente ou adaptada.

Gallica Bibliothèque nationale de France

Gallica é o nome da Biblioteca digital da Biblioteca Nacional da França, desenvolvida em colaboração com entidades cooperantes e estando disponível em linha desde 1997[6]. A atual versão da Biblioteca digital data de 2015 e engloba cerca de 5 milhões de números de jornais e de revistas[7].

O projeto subjacente à Gallica remonta a 1988, aquando do lançamento da nova Biblioteca nacional francesa, vindo a ganhar expressão com o desenvolvimento das tecnologias de informação e comunicação ocorrido na década de 90.

A partir da página inicial de acesso à imprensa e revistas, é possível percorrer um conjunto de categorias pré-estabelecidas por género, por local e por temática (ver figuras 7 e 8).

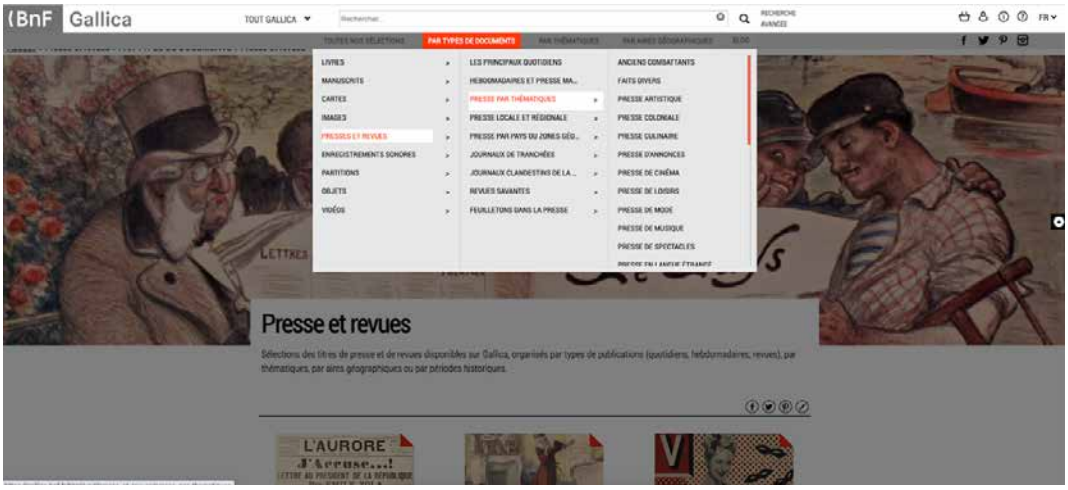

Figura 7
Exemplo 1 da apresentação da lista de temas
da imprensa e revistas. Fonte: gallica.bnf.fr / BnF.

5 Para mais informação sobre a linguagem LCHS, consultar: https://www.loc.gov/aba/publications/FreeLCSH/freelcsh.html#About
6 Disponível em: https://gallica.bnf.fr/html/und/presse-et-revues/presse-et-revues?mode=desktop
7 Ver, a este respeito, informação disponível em: https://gallica.bnf.fr/edit/und/a-propos

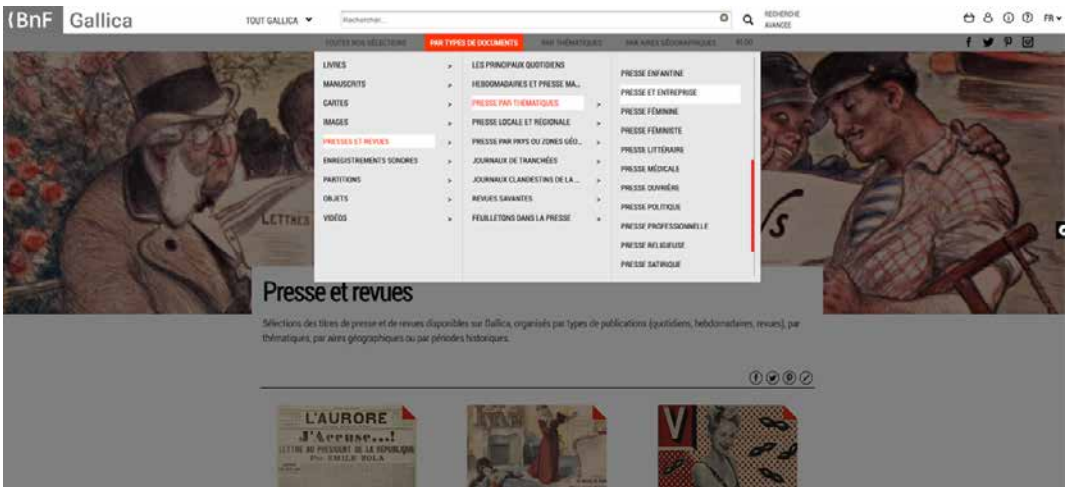

Figura 8
Exemplo 2 da apresentação da lista de temas
da imprensa e revistas. Fonte: gallica.bnf.fr / BnF.

A pesquisa por assunto em linguagem natural é realizada a partir do menu de pesquisa avançada comum a toda a Biblioteca digital, sendo necessário limitar a pesquisa à coleção de jornais e revistas. Neste menu, por outro lado, é oferecida uma lista de temas mais alargada do que a disponibilizada no acesso à imprensa e revistas (ver figuras 9, 10 e 11).

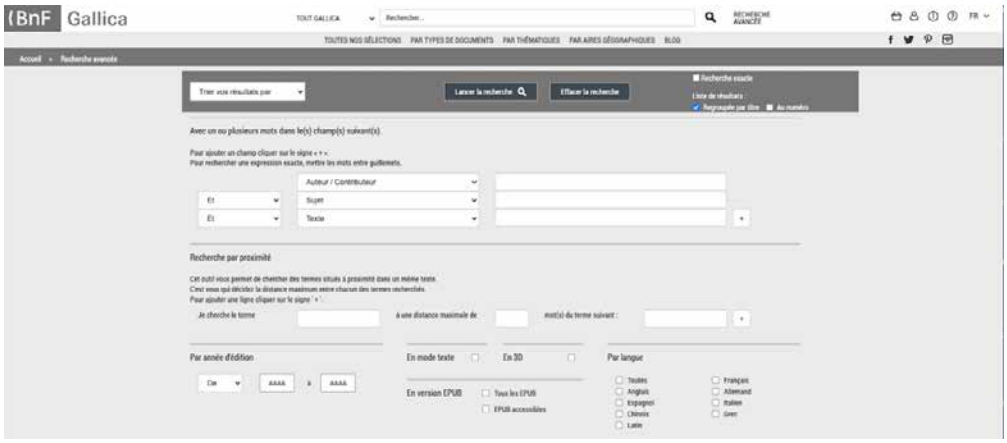

Figura 9
Imagem 1 do menu de pesquisa avançada.
Fonte: gallica.bnf.fr / BnF.

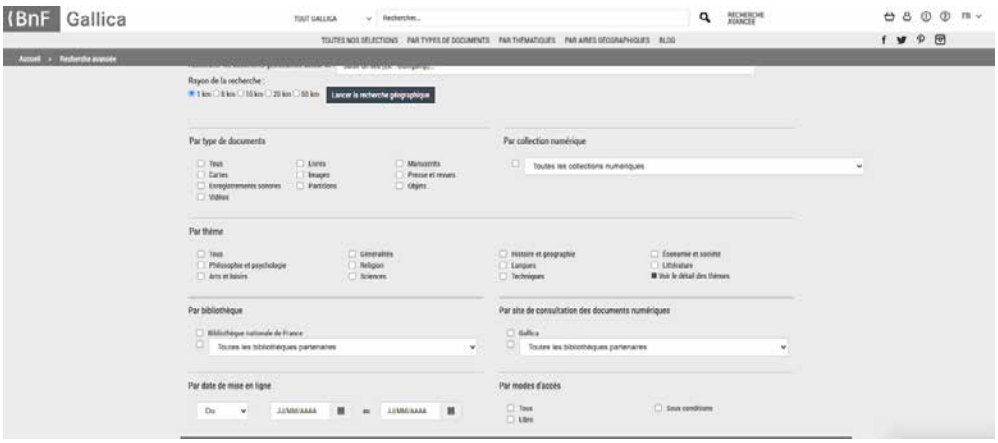

Figura 10
Imagem 2 do menu de pesquisa avançada.
Fonte: gallica.bnf.fr / BnF.

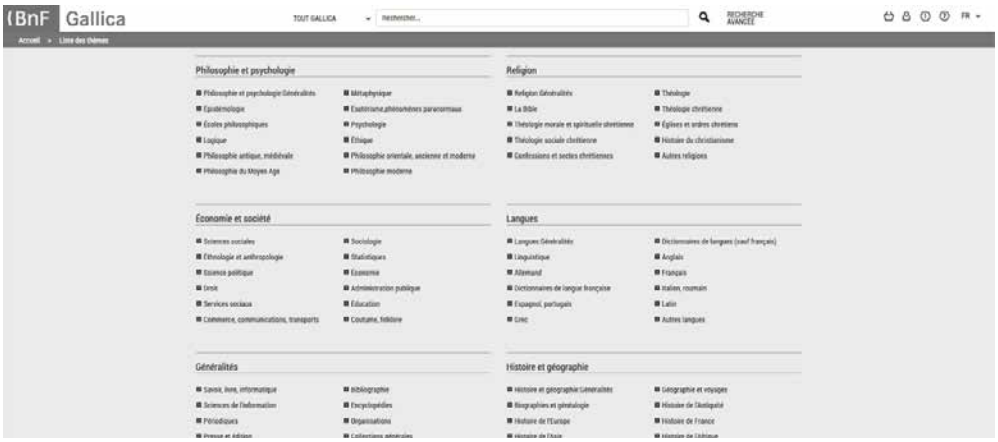

Figura 11
Exemplo da apresentação da lista completa
de temas da Gallica. Fonte: gallica.bnf.fr / BnF.

É de referir que a BnF utiliza a linguagem controlada Rameau na indexação por assuntos das suas coleções, desenvolvendo a partir desta linguagem categorias mais genéricas para a organização dos recursos digitalizados. O RAMEAU (Répertoire d'autorité-matière encyclopédique et alphabétique unifié) é a linguagem controlada de indexação por assuntos utilizada, também, pelas bibliotecas universitárias e bibliotecas públicas francesas. A linguagem Rameau tem sido desenvolvida desde 1980, em articulação com o Répertoire de vedettes-matière da Universidade Laval, Quebec, e com a lista de pontos de acesso por assuntos da Biblioteca do Congresso (Library of Congress Subject Headings)[8].

8 Ver, a este respeito, a informação disponível em: https://rameau.bnf.fr/

Os sistemas de organização do conhecimento e a indexação de jornais

A singularidade dos textos digitalizados de jornais requer uma abordagem por assunto em dois níveis. A abordagem que aqui se apresenta é aplicável quer ao tratamento técnico por assunto dos jornais, quer aos artigos dos jornais. Devemos, portanto, considerar a indexação do assunto geral de um determinado jornal, podendo incluir o nome geográfico, e devemos, também considerar, a indexação dos artigos. Ou seja, neste caso, pode envolver nome comum como assunto, nomes de pessoa, família e coletividades como assunto e, ainda, nomes geográficos. A complexidade do estabelecimento de pontos de acesso por assunto recorrendo ao controlo de autoridade, ou seja, a constituição de pontos de acesso controlados, tendo por base vocabulários controlados, é difícil, morosa e requer uma especial atenção à eventual utilização de sistemas de organização do conhecimento diferentes entre os diversos participantes, no caso de coleções com recursos de origens diversas.

O nível de profundidade do tratamento técnico fará uso das duas variáveis recorrentes no processo de indexação por assuntos, ou seja, a especificidade e a exaustividade. Entende-se por exaustividade a quantidade de conceitos retirados para caraterizar o conteúdo do texto e por especificidade exatidão aplicada na representação de um texto por um termo de indexação (Wellisch, 1996, pp. 175, 439). É necessário ser tão específico como o texto que estamos a tratar e utilizar termos mais genéricos ou mais específicos, de acordo com o aprofundamento que se exige. Percebe-se, portanto, que o tratamento dos artigos dos jornais requer uma maior especificidade e exaustividade na indexação por assunto, pois tal contribui para uma maior precisão na recuperação da informação. E, por este motivo, o fator consistência na seleção dos termos a utilizar é considerado como de especial relevo na indexação de notícias de jornais (Browne & Jermey, 2007, pp. 156-157).

Não é objetivo deste trabalho a discussão da fase de análise do conteúdo dos artigos, sendo que, na atualidade, é possível recorrer a sistemas automáticos ou semi-automáticos para o cumprimento desta tarefa. Porém, o desenvolvimento e aplicação dos sistemas de organização do conhecimento exige a presença humana na conceção das soluções mais adequadas a cada sistema a ser desenvolvido, não obstante a aplicação de técnicas automáticas para atribuição de termos de vocabulário controlado em elevados volumes de recursos em análise.

Chegados a este ponto, importa clarificar o que se entende por sistemas de organização do conhecimento. De acordo com Mazzochi (2018, p. 54), trata-se de um termo genérico usado para referir uma ampla gama de linguagens (por exemplo, tesauros, esquemas de classificação e ontologias), que foram concebidas para diferentes fins e em distintos momentos históricos, tendo em comum a função de apoiar a organização da informação e do conhecimento, de forma a facilitar a sua gestão e recuperação.

O uso das classificações tem sido praticado em bibliotecas na organização dos jornais, em lugar da tradicional organização por ordem alfabética e data (Kuhn, 1999, p. 106-113). Todavia, a sua aplicabilidade às bases de dados é limitada quando comparando com as linguagens vocabulares, como os tesauros e as listas estruturadas.

A utilização de linguagens vocabulares, como os tesauros e as listas estruturadas, tem sido frequente na indexação de jornais históricos selecionados (Barlow, 2009, pp.2-6; Hirst, 2013, pp. 158-162). Deste modo, as listas de índices são ajustadas à especificidade do jornal em questão e vão de encontro às necessidades dos utilizadores. Apesar das recomendações presentes em livros de reconhecida autoridade relativamente à sua utilização, também se encontram vozes discordantes ao sugerir que as listas de pontos de acesso por assuntos normalizadas e habitualmente em uso nas Bibliotecas não devem ser utilizadas e sim criadas listas para a cobertura temática da especificidade dos jornais a considerar (Wellisch, 1996, p. 326-327; Cleveland e Cleveland, 2013, pp.233-236).

Ou seja, se considerarmos a indexação por assunto dos jornais, é aplicável uma linguagem como a Library of Congress Subject Headings ou a Rameau acima identificadas. Mas, se se pretender proceder ao estabelecimento de pontos de acesso para artigos dos jornais, as referidas linguagens podem não ser suficientemente específicas para representar o conteúdo em causa.

Os sistemas de organização do conhecimento em bases de dados de jornais digitalizados em Portugal

Embora sendo de menor dimensão face aos exemplos de bases de dados internacionais acima descritos, é imperativo referir duas bases de dados portuguesas de jornais digitalizados, pela sua importância para o assunto em epígrafe. Trata-se da Biblioteca Nacional Digital, coleção de Periódicos, e da Hemeroteca Digital, desenvolvida pela Hemeroteca Municipal de Lisboa.

A Biblioteca Nacional Digital (BND)[9], desenvolvida pela Biblioteca Nacional de Portugal, tem como objetivo oferecer o acesso em linha, universal e gratuito, a conteúdos digitalizados de manuscritos e impressos, entre os quais se encontram as publicações periódicas. Todos os conteúdos da BND têm atribuído um estatuto de direitos. No caso de obras já caídas em domínio público, os conteúdos são disponibilizados com a marca Public Domain e podem ser usados livremente. A BND permite o acesso a um vasto conjunto de títulos desde o século XVII, mas nem todos os títulos têm associado o acesso a cópia pública. É possível percorrer os

9 Disponível em: https://bndigital.bnportugal.gov.pt/

jornais por século ou por título, não estando disponibilizado o acesso por assuntos (ver figura 12). Após ser selecionada uma dada publicação, pode ser visualizado o correspondente registo no catálogo da Biblioteca Nacional de Portugal e observada a indexação do jornal, quando está disponível (ver figuras 13 e 14).

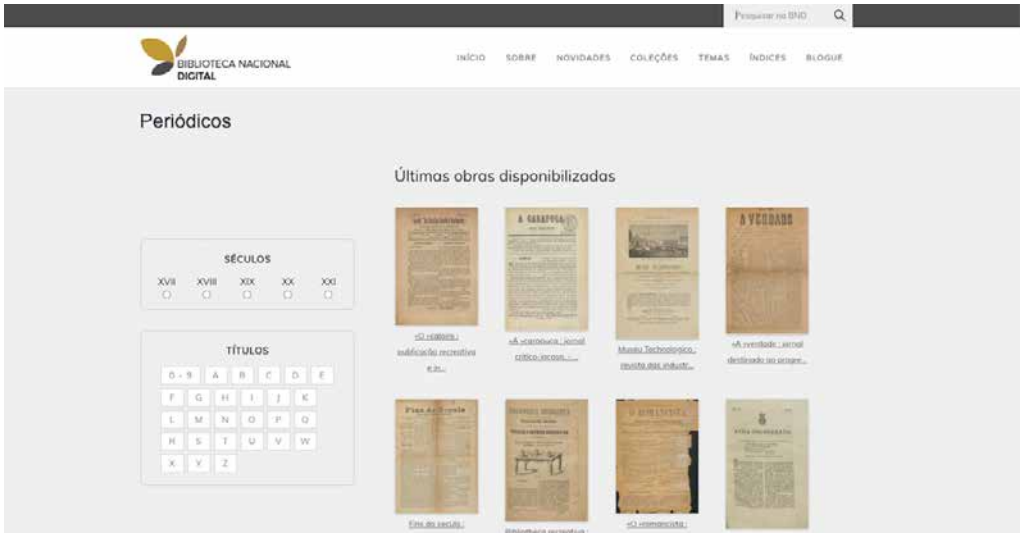

Figura 12
Página da BND relativa aos periódicos. Fonte: Biblioteca Nacional de Portugal.

Figura 13
Visualização do registo do jornal na BND. Fonte: Biblioteca Nacional de Portugal.

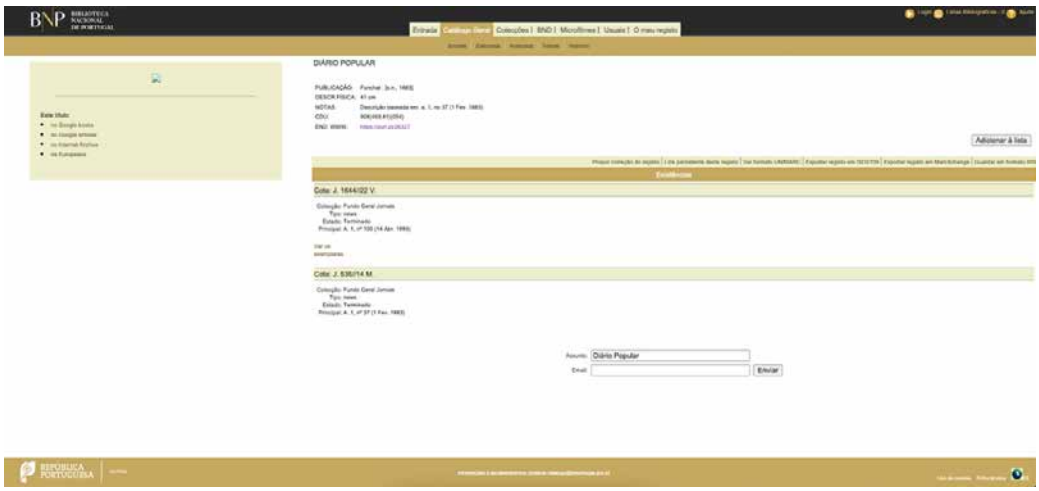

Figura 14
Visualização do registo do jornal no catálogo da Biblioteca Nacional de Portugal. Fonte: Biblioteca Nacional de Portugal.

Observando a figura 14, é possível verificar a existência da atribuição da classificação CDU ao jornal catalogado. A Classificação Decimal Universal é a classificação utilizada nas bibliotecas portuguesas para a organização dos catálogos e das coleções em livre acesso. É uma classificação assente numa estrutura predominantemente hierárquica, permitindo a atribuição de subdivisões, por exemplo, de forma, de lugar e de tempo[10].

Apesar de não estar disponível a indexação recorrendo a linguagem vocabular, é de referir que a Biblioteca Nacional de Portugal editou em 2003 um manual para indexação que veicula um conjunto de instruções para a determinação da forma dos termos (Portugal, 2003). Este manual é de significativa utilidade para a construção de pontos de acesso por assunto numa base de dados de jornais digitalizados, aplicando-se os seus princípios à indexação de artigos de jornais. Apesar de existir normalização internacional que determina as orientações para a construção de tesauros para a recuperação da informação, mais concretamente a norma ISO 25964-1:2011, o manual esclarece dúvidas específicas da adaptação dos princípios internacionais à particularidade da língua portuguesa.

O outro exemplo de base de dados que disponibiliza o acesso a coleções de jornais digitalizados é, conforme referimos, a Hemeroteca Digital[11]. Permitindo o acesso a jornais e revistas de acesso livre por estarem em domínio público, disponibiliza a consulta a listagens de índices de títulos, autores, acesso cronológico, nome geográfico e por géneros de imprensa (ver figuras 15 e 16).

10 Uma versão reduzida da CDU está disponível para consulta em: http://www.udcsummary.info/php/index.php?lang=pt
11 Disponível em: http://hemerotecadigital.cm-lisboa.pt/index.htm

Figura 15
Imagem da página principal da Hemeroteca Digital. Fonte: Hemeroteca Digital.

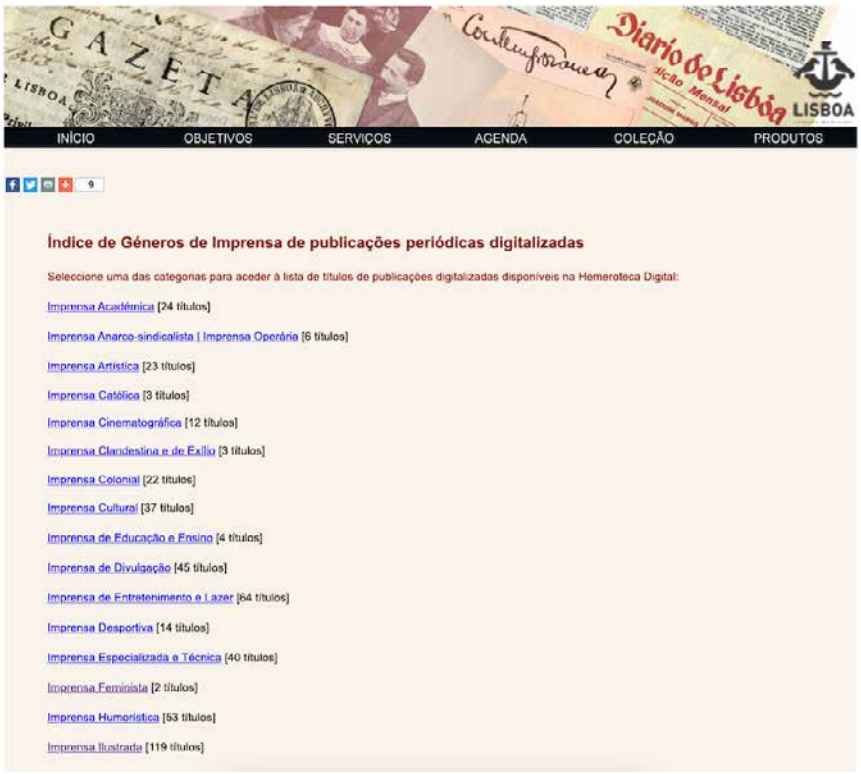

Figura 16
Imagem do índice de géneros de imprensa de publicações periódicas digitalizadas. Fonte: Hemeroteca Digital.

Conclusão

Encerramos este trabalho da forma como o iniciámos, ou seja, sublinhando a importância do acesso aos jornais históricos para o estudo da história do jornalismo. O percurso traçado neste estudo permite concluir que existem acentuadas assimetrias nas prioridades e no investimento que os países conferem a este propósito. Projetos como a Biblioteca Nacional Digital e a Hemeroteca Digital constituem um princípio do que deveria ser uma política de digitalização e indexação de jornais históricos a nível nacional. No entanto, as modalidades de pesquisa oferecidas e a organização por assunto dos títulos e artigos disponibilizados está muito aquém dos exemplos internacionais, exigindo, portanto, um considerável investimento em meios humanos e tecnológicos.

A utilização de sistemas de organização do conhecimento no tratamento de títulos de jornais e de artigos, em prática nos exemplos de bases de dados internacionais descritas no presente trabalho, enriquece as modalidades de pesquisa e permite o acesso por assunto, ultrapassando as contingências decorrentes da leitura integral dos textos das publicações selecionadas para análise por parte de investigadores, sobretudo quando se trata de estudar um corpus de análise de elevada dimensão.

O estudo aqui apresentado não se encerra nestas linhas. Para explorar outras formas de organização de conhecimento aplicadas em bases de dados relativas a coleções de jornais históricos digitalizados, justifica-se um estudo aprofundado de bases de cariz comercial, pois o investimento de que são alvo, certamente, se reflete na oferta de funcionalidades de pesquisa. Por outro lado, a quase inexistente literatura sobre as políticas de indexação por assuntos das coleções de jornais levadas a cabo em Serviços de informação como Arquivos e Bibliotecas, abre, também, a possibilidade à exploração de técnicas e práticas que podem revelar-se úteis no contexto da construção de sistemas de acesso por assunto aos jornais históricos. Esperamos, com o presente trabalho, contribuir para o estímulo ao desenvolvimento de outros estudos sob diferentes perspetivas e utilizando outras metodologias.

Referências bibliográficas

Barlow, C. (2009). Serials indexing: from journals to databases. *The Indexer*, *27*(1), 2-6. https://doi.org/10.3828/indexer.2009.2

Birkner, T., Koenen, E., & Schwarzenegger, C. (2018). A Century of Journalism History as Challenge. *Digital Journalism*, *6*(9), 1121-1135. https://doi.org/10.1080/21670811.2018.1514271

Browne, G. & Jermey, J. (2007). *The indexing companion*. Cambridge University Press.

Cleveland, D. B., & Cleveland, A D. (2013). *Introduction to indexing and abstracting* (4th ed.). Libraries Unlimited.

Europeana Newspapers. (2015). *Europeana Newspapers. Special issue – Final Report*. https://europeananewspapers.github.io/?page=1

Gabriele, S. (2014). Transfiguring the Newspaper. *Amodern 2: Network Archaeology*. htttp://amodern.net/article/transfiguring-the-newspaper.pdf

Gabriele, S. (2014). Transfiguring the Newspaper. *Amodern 2: Network Archaeology*. htttp://amodern.net/article/transfiguring-the-newspaper.pdf

Hansen, K. A. & Paul, N. (2015, April, 15-16)). *News Archive Chaos: A Case Study*[Paper presentation]. IFLA News Media and Audiovisual and Multimedia Sections' Conference "Transformation of the online news media: implications for preservation and access", Stockholm, Sweden, National Library of Sweden. https://www.ifla.org/files/assets/newspapers/Sweden_2015/6_-_hansen_and_paul_ifla_2015_news_archive_chaos.pdf

Hirst, J. (2013). On indexing the Argus. *The Indexer*, *31*(4), 158-162. https://doi.org/10.3828/indexer.2013.51

International Organization for Standardization. (2011). *Information and documentation — Thesauri and interoperability with other vocabularies* (ISO Standard No. 25964-1:2011). https://www.iso.org/standard/53657.html

Kuhn, T. (1999). Classifying Newspapers Using Dewey Decimal Classification. *Library Resources & Technical Services*, *43*(2), 106-113. https://doi.org/10.5860/lrts.43n2.106

Mazzocchi, F. (2018). Knowledge organization system (KOS). *Knowledge Organization*, *45*(1), 54-78. https://doi.org/10.5771/0943-7444-2018-1-54

Mouhot, J. (2010). Archival Review: ProQuest Historical Newspapers, *Contemporary British History*, *24*(1), 131-134. https://doi.org/10.1080/13619460903553867

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

Portugal. Biblioteca Nacional de Portugal. (2001). *Indexação: terminologia e controlo de autoridades (manual)*. Biblioteca Nacional.

Tartaglia, S. (2004). Authority Control and Subject Indexing Languages. In Arlene G. Taylor & Barbara B. Tillet, (Eds.), *Authority Control in Organizing and Accessing Information: Definition and International Experience* (pp.). The Haworth Information Press. https://doi.org/10.1300/J104v38n03_04 23

Wellisch, H. H. (1996). *Indexing from A to Z* (2nd ed.). H.W. Wilson.

Willems, M. & Atanassova, R. (2015). Europeana Newspapers: searching digitized historical newspapers from 23 European countries. *Insights*, *28*(1), 51-56. http://doi.org/10.1629/uksg.218