Margarida Isabel Abrantes Neves

Construção de Hemerotecas Digitais: Uma Proposta de Modelo

Dissertação de Mestrado em Ciência da Informação, orientada pela Doutora Maria Manuel Borges e coorientada pelo Dr. António Tavares Lopes, apresentada ao Departamento de Filosofia, Comunicação e Informação da Faculdade de Letras da Universidade de Coimbra

2018

· U C ·

Faculdade de Letras

Construção de Hemerotecas Digitais: Uma Proposta de Modelo

Ficha Técnica:

Tipo de trabalho	**Trabalho de Projeto**
Título	**Construção de Hemerotecas Digitais: Uma Proposta de Modelo**
Autora	**Margarida Isabel Abrantes Neves**
Orientadora	**Professora Doutora Maria Manuel Borges**
Coorientador	**Dr. António Tavares Lopes**
Júri	**Presidente do Júri: Doutora Maria da Graça Melo Simões** **Dr. António Fernando Tavares Lopes** **Doutora Maria Cristina Gonçalves Guardado**
Identificação do Curso	**Mestrado em Ciência da Informação**
Área científica	**Ciência da Informação**
Data da Defesa	**26-10-2018**
Nota	**18 Valores**
Imagem da Capa	**https://www.flickr.com/photos**

· U C ·

UNIVERSIDADE DE COIMBRA

Para ser grande, sê inteiro: nada
Teu exagera ou exclui,
Sê todo em cada coisa. Põe quanto és
No mínimo que fazes.
Assim em cada lago a lua toda
Brilha, porque alta vive.
(Ricardo Reis, 1933: 148)

SUMÁRIO

AGRADECIMENTOS

À minha Orientadora, a Professora Doutora Maria Manuel Borges deixo aqui as minhas palavras de agradecimento, por me acompanhar desde a Licenciatura e pela disponibilidade de me acompanhar em todo este processo. Pela sua incrível inteligência, dedicação e contribuição para a Ciência da Informação. Expresso a minha grande admiração por todo o seu trabalho e capacidade.

Ao meu coorientador, o Dr. António Tavares Lopes por de igual modo me acompanhar desde a Licenciatura e por toda a amabilidade e disponibilidade. Por ter apostado em mim quando escolhi esta temática.

A todos os meus Professores da FLUC, por toda a partilha de conhecimento, não só na área de estudo, como também a nível de enriquecimento pessoal. Em especial, à Professora Doutora Maria da Graça Melo Simões pelo carinho.

Agradecer é demonstrar carinho por aqueles que estiverem sempre no meu caminho, nas derrotas, nas conquistas e nas vitórias sem que nada me pedissem em troca.

Em primeiro lugar, agradeço à fé que sempre tive e que nunca deixou que a luz se apagasse nas caminhadas mais difíceis.

De um modo particular deixo as minhas palavras:

Aos meus pais, Aires e Mila. Os meus pilares. A força e o carinho de mãos dadas. Pelos abraços, palavras sábias e pelo colo.

Ao meu querido irmão Rodrigo e à sua alegria, energia e sabedoria.

À minha querida avó, Aurorita, por todas as orações, por me aquecer o coração e por me enxugar as lágrimas.

Aos meus queridos mais pequeninos, os meus primos Rafaela e Francisco, por todos os miminhos e carinhos.

À Ana Luísa e ao seu incrível desempenho na Biblioteca Municipal Eduardo Lourenço (Guarda). A minha fonte de inspiração no que toca à Ciência da Informação, ao esforço e dedicação da profissão. Que me ensinou a nunca desistir dos nossos sonhos.

À BMEL por me acolher pela segunda vez. Por me deixar explorar os caminhos que tanto quero percorrer no futuro.

Às flores do meu jardim: Carol e Magda que estiveram sempre comigo. O motivo da minha força e inspiração. Por todas as conversas, conselhos, sorrisos, choros e segredos. Por serem as minhas melhores amigas.

Às minhas colegas de curso, Aida, Maria, Coxixo e Verónica que me acompanham há cinco anos e por quem nutro um grande carinho e amizade.

Ao Rodrigo, por todas as palavras de carinho, atenção e força. Por nunca me ter deixado baixar os braços e manter sempre o rumo deste barco. Por todo o amor!

À Inês, à Mafalda e à Rita, família que Coimbra me deu. Por mais que a distância nos separe, o carinho será sempre maior.

À Raquel e à Rute. Pelas experiências partilhadas. Pelos sorrisos e abraços.

À Tânia, que apareceu em último lugar, mas que ocupa grande parte do meu coração. Por me ensinar a ser profissional, trabalhadora, perspicaz e inteligente. Por ser carinhosamente preocupada e pela sua incrível capacidade de me ajudar em todas as circunstâncias, no escritório e em todo o processo de elaboração deste projeto.

RESUMO

A digitalização de jornais é uma prática cada vez mais recorrente entre as bibliotecas e as instituições que armazenam acervos com o objetivo de organizar, preservar a memória de uma cidade e/ou pais, facilitar o acesso à informação, salvaguardar o património e proteger contra possíveis perdas. A conceção de uma hemeroteca digital deve ser vista como uma ferramenta essencial de forma a proporcionar o acesso à informação e é implícito nas questões de preservação e disseminação das coleções. O presente trabalho tem como objetivo propor um guia de boas práticas que sumaria um conjunto de orientações para a criação de uma hemeroteca digital. A metodologia de trabalho compreendeu fases distintas: (i) revisão da literatura sobre os conceitos fundamentais para a descrição, (ii) análise dos projetos National Digital Newspaper Program, a Europeana e o Making of America visando a recolha de informação sobre planos de digitalização e publicação e (iii) a aplicabilidade desses requisitos essenciais transformados numa proposta para a construção de uma hemeroteca digital.

PALAVRAS-CHAVE: Hemeroteca Digital; Plano de Digitalização; Acesso à Informação; NDNP; Europeana; MoA.

ABSTRACT

The digitization of newspapers is an increasingly recurrent practice among libraries and institutions that store collections with the aim of organizing, preserving the memory of a city and / or country, facilitating access to information, safeguarding heritage and protecting against possible losses. The design of a digital newspaper library must be seen as an essential tool in order to provide access to information and is implicit in the issues of preservation and dissemination of collections.

The present work aims to propose a guide to good practices that would summarize a set of guidelines for the creation of a digital newspaper library. The work methodology comprised distinct phases: (i) review of the literature on the fundamental concepts for the description, (ii) analysis of the National Digital Newspaper Program, Europeana and Making of America projects aiming at collecting information on digitization plans and and (iii) the applicability of these assumptions transformed into a proposal for the construction of a digital newspaper archive.

KEY WORDS: Digital Newspaper Library; Digitization Plan; Information Access; NDNP; Europeana; MoA.

SIGLAS E ABREVIATURAS

ASCII – American Standard Code Information Interchange

CBSR -The Centre of Bibliographical Studies and Research

CDNC – California Newspaper Collection

CGI – Common Gateway Interface

CMSI – Cimeira Mundial Sobre a Sociedade da Informação

DPI- Ponto de Polegada

GIF - Graphic Interchange Format

HTI – Humanities Text Initiative

ICDAR – International Conference on Document Analysis and Recognition

JPEG – Joint Photographic Experts Groups

MoA – Making of America

MUFI – Medical Unicode Font Initiative

NCSR – National Centre for Scientific Research

NDNP – National Digital Newspaper Program

NER – Named-entity Recognition

NEH – National Endowment for the Humanities

NMI – Non-maskable Interrupt

OCR – Optical Character Recognition

OLR – Optical Layout Recognition

PAGE – Page analysis and Ground Truth Elements

PDF – Portable Document Format

PNG – Portable Network Graphics

RLG – Research Libraries Group

SGML – Standard Generalized Markup Language

TEI – Text Encoding Initiative

TIC – Tecnologias de Informação e Comunicação

TIFF – Tagged Image File Format

UIT – União Internacional das Telecomunicações

XML – Extensible Markup Language

LISTA DE FIGURAS E QUADROS

INTRODUÇÃO

O papel das bibliotecas públicas é disponibilizar recursos que sejam de interesse para os seus utilizadores, sem que haja distinção da condição social, crença, raça ou nacionalidade. Achamos assim pertinente reunir práticas com um conjunto de orientações para a criação de uma biblioteca digital de uma coleção de jornais uma vez que são instituições carentes de projetos deste género. As bibliotecas públicas são as instituições responsáveis por armazenar, conservar e difundir acervos tão importantes como os jornais da comunidade onde estão inseridas, de forma a enaltecer o valor do património bibliográfico e do património cultural.

As bibliotecas digitais armazenam materiais em formato eletrónico de maneira a serem acedidos mais facilmente, prática e ao alcance de todos (*Web*), por diferentes tipos de utilizadores e de distintos lugares. É sabido que uma das fontes mais importantes, como parte do conhecimento cultural, foi e ainda é armazenada nos jornais. Este poderoso meio de comunicação teve o seu nascimento com a imprensa de Gutenberg no séc. XV, tirando a cultura dos círculos minoritários e disponibilizando-a nos círculos maioritários, colocando-a ao alcance de todos.

Poucas fontes históricas permitem tantas possibilidades de investigação para o historiador, investigador ou um simples curioso, quanto o jornal. Estamos assim a referir-nos 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 registar os mais distintos fenómenos culturais, políticos, económicos, sociais e até mesmo naturais. (Carvalho, 2016).

O presente trabalho de projeto tem como principal objetivo reunir um conjunto de requisitos funcionais de forma a criar um guia de boas práticas para a criação de uma Biblioteca Digital de Jornais, isto é, uma Hemeroteca Digital. Este conjunto de orientações é representado através de soluções fiáveis de forma a combater a problemática de todo o processo de preservação e recuperação de informação (de jornais) numa Biblioteca Pública: desde a digitalização até à sua disponibilização na *web*. Todo este processo é tido como importante no que concerne à facilidade do acesso às coleções, na divulgação, valorização e preservação do património histórico-cultural de uma determinada instituição. A área das bibliotecas digitais tem atraído diversos

pesquisadores nos últimos anos; foram iniciados muitos projetos interessantes, lidando com diversas questões ainda em aberto sobre a função e o objetivo para a criação de bibliotecas digitais. As TIC provocaram uma revolução na forma e nos métodos de gestão, armazenamento, processamento e difusão da informação. A grande mudança foi provocada pela renovação do texto em suporte papel para o suporte eletrónico. O contexto digital é real, entretanto, significativa parcela da população não possui acesso aos seus recursos, sendo a exclusão digital também uma realidade (Nonato, Borges, Maculan, & Lima, 2008).

A metodologia de trabalho compreendeu fases distintas como, a reunião de conceitos importantes para identificação do problema, recolha de informação sobre planos de digitalização e publicação e a transformação dos requisitos funcionais encontrados nos três projetos já referidos, numa tabela dividida em requisitos obrigatórios, recomendados e facultativos, que correspondam às fases de digitalização, processamento e publicação. A pesquisa foi feita em bases de dados e repositórios científicos como o RCAAP, a B-On,o EBSCOhost e o IBICT com os termos: "Biblioteca Digital de Jornais"; "Hemeroteca Digital" e "Planos de Digitalização de Jornais".

Esquema 1: Metodologia de Trabalho.

O trabalho encontra-se dividido em três capítulos. O primeiro capítulo consiste numa revisão de literatura com base numa análise crítica dos estudos e trabalhos que vão de encontro à temática das bibliotecas digitais de jornais, métodos de digitalização e a sua aplicação. Optou-se por escolher as bibliotecas públicas para uma possível aplicação dos métodos de digitalização e publicação, que foram estudados a partir de distintos projetos que se revelassem transversais e facilmente adaptáveis às circunstâncias de cada instituição: a capacidade financeira e de recursos humanos. No segundo capítulo foram analisados três distintos projetos de digitalização, um europeu e dois americanos. Os projetos aqui identificados foram selecionados devido às caraterísticas heterogéneas das suas coleções, são eles: o *National Digital Newspaper Program*, um arquivo digital de

jornais criado entre a Biblioteca do Congresso e o *National Endowment for Humanities* que os disponibiliza numa plataforma que inclui informações históricas contextualmente relevantes; a Europeana, uma biblioteca virtual desenvolvida por países da União Europeia com o objetivo de partilhar imagens, documentos, mapas e jornais de forma a disseminar a cultura e história europeias. São permitidas pesquisas por nome, por tipo de documento e ainda por nomes de localidades; e o *Making of America*, uma colaboração entre a Universidade de Cornell e a Universidade de Michigan, nos EUA que oferece uma navegação de pesquisa de tipo booleana, bibliográfica e simples.

Estes três projetos estudados foram desenhados essencialmente para responder a diversos suportes de informação, não estando unicamente indicados para o livro impresso e/ou para suporte em microfilme, como também estão apropriados para publicações periódicas (jornais). Os projetos são reconhecidos a nível científico e global uma vez que nos apresentam, de forma íntegra, todo o processo: as principais regras para a digitalização, a gestão dos documentos que pretendemos digitalizar, a escolha dos *Softwares* e *Hardwares*, o formato recomendado (Dpi), o Reconhecimento Ótico de Carateres e a transformação da informação em metadados.

A escolha destes projetos como base para a orientação para o desenho de projeto de uma Hemeroteca Digital foi uma consequência da facilidade com que conseguimos aceder a toda a informação disponibilizada, visto que há uma grande reprodução de literatura e consequentemente uma vasta produção de bibliografia, disponível em linha. Aqui, são tecidos apenas alguns requisitos funcionais que poderão servir como base numa instituição pública e que podem ser enriquecidos se se optar por uma consulta integral dos mesmos, de forma a atingir os objetivos pretendidos.

Após a reunião de aspetos pertinentes como o processo desde a seleção dos documentos a digitalizar, o seu tratamento e a digitalização, à aplicação do OCR e a transformação em metadados, foram identificadas e sumariadas num quadro os requisitos funcionais para a criação de uma Hemeroteca Digital. Além disso, os requisitos encontram-se categorizados em elementos obrigatórios, recomendados e facultativos em cada um dos campos a que se refere, nomeadamente, no campo da Digitalização, no campo do Processamento e no campo da Publicação.

CAP. 1 AS HEMEROTECAS DIGITAIS NAS BIBLIOTECAS PÚBLICAS

1.1 As funções das Bibliotecas Públicas

O Manifesto da IFLA/UNESCO para as Bibliotecas Públicas (Manifesto, 1994)[1] define a biblioteca pública – porta de acesso local ao conhecimento – fornece as condições básicas para uma aprendizagem contínua, para uma tomada de decisão independente e para o desenvolvimento cultural dos indivíduos e dos grupos sociais. (...) A biblioteca pública é o centro local de informação, tornando prontamente acessíveis aos seus utilizadores o conhecimento e a informação de todos os géneros (Manifesto, 1994). Os serviços da biblioteca pública devem ser oferecidos com base na igualdade de acesso para todos, sem distinção de idade, raça, sexo, religião, nacionalidade, língua ou condição social. Todos os grupos etários devem encontrar documentos adequados às suas necessidades. As coleções e serviços devem incluir todos os tipos de suporte e tecnologias modernas apropriadas assim como materiais tradicionais. É essencial que sejam de elevada qualidade e adequadas às necessidades e condições locais. As coleções devem refletir as tendências atuais e a evolução da sociedade, bem como a memória da humanidade e o produto da sua imaginação. Muito mais do que definir o público alvo ou tipos de acervos, o Manifesto "proclama a confiança que a Unesco deposita na Biblioteca Pública, enquanto força viva para a educação, a cultura e a informação, e como agente essencial para a promoção da paz e do bem-estar espiritual nas mentes dos homens e das mulheres" (IFLA, 1994). As definições e diretrizes contidas neste manifesto têm servido, desde 1994, como parâmetro para as bibliotecas públicas de todo o mundo.

Nem só como um depósito de livros se define uma Biblioteca pública, esta, tem hoje um papel fundamental na sociedade, na medida em que se torna um local de interação, debates e manifestações culturais e artísticas, extrapolando o seu papel de democratização da cultura. Como afirma Ferraz (2014), é um centro de promoção cultural, atuando como veículo para o exercício da cidadania. Desempenha um papel fundamental no que concerne à democratização do acesso à informação, na medida em que se recebe, sem

[1] Estabelece o conceito de biblioteca e as suas missões-chave. Proclama a confiança que a UNESCO deposita na biblioteca pública enquanto fator de promoção da educação, cultura e informação dos cidadãos. Foi elaborado pela Secção de bibliotecas públicas da IFLA e aprovado pela UNESCO em 1994.

distinção, qualquer pessoa independentemente da sua classe social, sexo, orientação sexual ou religião, tornando-se a mais democrática de todos os tipos de biblioteca.

A importância social da biblioteca pública está justamente em conseguir pensar nas necessidades da comunidade na qual ela está inserida, e saber reconhecer os interesses da população. Indo um pouco mais além, deve-se pensar na potencial procura, ainda não reconhecida pela população, mas que deve ser oferecida e incorporada às políticas culturais vigentes.

O Manifesto da IFLA/UNESCO para as Bibliotecas Públicas, lançado em 1994, tem servido de parâmetro para pensar no conceito das bibliotecas públicas, na sua missão e no seu papel social. Com o objetivo de ampliar as conceções de atuação das bibliotecas públicas, a IFLA lança o Manifesto sobre transparência, no combate à corrupção, na boa gestão por parte das entidades governativas, reafirmando o papel social da biblioteca pública.

A IFLA afirma que:

> As bibliotecas são na sua verdadeira essência instituições transparentes, dedicadas a colocar à disposição de cada um e de todos as informações educacionais, científicas, técnicas e socialmente mais relevantes, mais acuradas e imparciais. Os materiais de informação e acessos providos pelas bibliotecas e os serviços de informação contribuem para o bom governo aumentando o conhecimento dos cidadãos e enriquecendo suas discussões e debates. As bibliotecas e os serviços de informação devem ampliar sua missão de modo a se tornarem componentes mais ativos do bom governo e na luta contra a corrupção. Em particular eles podem desempenhar um papel significativo informando aos cidadãos sobre seus direitos e garantias (IFLA, 2008).

1.2 Definição e Serviços das Bibliotecas Digitais

A partir da década de 1990, com o crescente domínio das Tecnologias da Informação e Comunicação (TIC), os bibliotecários iniciaram um processo em relação aos procedimentos automatizados, de forma a experimentarem a tecnologia como forma de representar um suporte ao serviço da biblioteca. Com a chegada das bibliotecas digitais, esta revolução tecnológica apresenta novas oportunidades de mudança. A criação de plataformas *web* com certas características e organização, denominadas de diferentes formas: "bibliotecas eletrónicas", "bibliotecas digitais" ou "bibliotecas virtuais" (Toutain, et al, 2005), é uma das alternativas para garantir que a literatura científica seja disseminada e alcance os utilizadores com a qualidade exigida. Acredita-se que a biblioteca digital é uma evolução da biblioteca tradicional, e supostamente esse fato dá-se a partir da década

de 1960 com o processo de informatização das bibliotecas(Assunção, 2011, p. 2). É importante ter em conta que as bibliotecas possuem uma longa e complexa história de mudanças tecnológicas, e com o advento da internet, a biblioteca digital desempenhou um papel fundamental na comunicação científica (Assunção, 2011). A biblioteca digital surge num contexto em que se sobrepõe a necessidade de guardar, organizar e disseminar toda a informação e o conhecimento produzidos pela humanidade no decorrer do tempo.

> Poderíamos atribuir também o nascimento da biblioteca digital como um sonho da biblioteca universal, que seria capaz de reunir todo o conhecimento, experiência e literatura humana para que não se perdessem com o tempo, assim como disseminar essa informação de maneira global (Assunção, 2011).

O conceito de biblioteca digital não é apenas o equivalente de repertórios digitalizados com o método de gestão da informação; é também um ambiente em que as coleções, os serviços e o pessoal se encontram.

Crespo, (2001), p. 7 *apud* Oppenheim, (1999) define biblioteca digital como:

> Un servicio de información en el cual todos los recursos informativos están disponibles en formato manipulable por computadoras y las funciones de adquisición, almacenamiento, preservación, recuperación, acceso y presentación del documento se llevan a cabo mediante el empleo de tecnología digital.

Segundo Fernández (2009, p.4) as bibliotecas digitais são organizações que promovem recursos, incluindo pessoal especializado para selecionar, estruturar, oferecer, traduzir, distribuir e preservar a integridade e garantir a permanência das coleções digitais, para que estejam disponíveis para uma ou várias comunidades. Este é um momento de avançar para além dos aspetos relativos à aquisição e ao processamento dos materiais, integrando fontes e materiais eletrónicos nos acervos e serviços. A biblioteca digital pode conter diferentes tipos de suportes informacionais sobre os mais variados assuntos. Assim, por exemplo, encontra-se todo o conteúdo informacional num único formato equivalente - o digital. A ênfase do foco da biblioteca digital é maior no acesso à informação e menor relativamente à coleção. Assim, a organização da biblioteca digital deve refletir-se nos documentos que os utilizadores necessitam, e não naqueles que o bibliotecário tem condições de incorporar ao acervo (Cunha, 2008).

Para (De Lima, De Oliveira, & De Santana, 2013) *apud* Silva e Garcia (2005) e Sayão (2008), as bibliotecas digitais tiveram a sua génese em Paul Otlet, com o sonho de uma biblioteca universal, Vanevar Bush, com a sua máquina amplificadora de memória e Theodore Holm Nelson com o projeto Xanadu e a sua representação do pensamento

associativo. Posteriormente com Tim Berners Lee idealizando e criando o sistema *WWW* (*World Wide Web)* para reunir virtualmente informações. Já (Pontes & Lima, 2012) compartilham o entender de que a biblioteca digital possui os mesmos objetivos da biblioteca tradicional e que deveria basear-se nos mesmos princípios, teorias e técnicas desenvolvidas pelo campo da Biblioteconomia e Ciência da Informação.

A biblioteca digital combina a estrutura e coleta da informação, tradicionalmente usada por bibliotecas e arquivos, com o uso da representação digital tornada possível pela informática. A informação digital pode ser acedida de forma rápida em todo o mundo, copiada para preservação, armazenada e recuperada. Uma biblioteca digital – como uma coleção de informação digitalizada e organizada – tem um potencial informacional que dificilmente terá sido alcançado por alguma biblioteca convencional, isto é, ela pode entregar a informação diretamente ao utilizador, possuindo a capacidade de executar estratégias de pesquisa de palavras isoladas ou expressões inteiras e o seu conteúdo informacional – seja ele em forma textual, sonora ou em imagens – não está exposto ao uso intensivo de um documento impresso (Cunha, 2008, p. 5).

O conceito de biblioteca digital constitui um subconjunto de um conceito mais extenso de biblioteca e não um substituto. Lima, Sousa e Dias (2015, p. 19) *apud* Sayão (2008, p.8) "desde início da computação que ficou claro que a automatização das bibliotecas traria um extraordinário ganho de produtividade aos processos biblioteconómicos por conta da natureza e do volume de dados tratados pelas bibliotecas". Assim, o uso da informática nas bibliotecas origina uma prática biblioteconómica que substitui a criação de catálogos por portais de acesso, integrando o armazenamento, a consulta e suprimento em formato dos próprios documentos legível na sua diversidade. Lima, Sousa e Dias (2015, p. 19) *apud* Le Crosnier (2005) afirma que para a *Association of Research Libraries*[2] "as coleções de bibliotecas digitais não se contentam com referências, mas se interessam por todos os artefactos digitais que não podem ser apresentados ou representados de forma impressa".

Nos conceitos encontrados na Ciência da Informação o entendimento acerca de "biblioteca digital" contém representações digitais dos objetos e deve ser acessível através da internet, embora não para todos. Mas a ideia da digitalização é a única caraterística de uma biblioteca digital em que há um consenso. Autores como (Lima, Souza, & Dias, 2015) compreendem

[2] *Association of Research Libraries* é uma organização sem fins lucrativos de 125 bibliotecas de pesquisa no Canadá e nos Estados Unidos da América que compartilham missões de pesquisa. Disponível em: http://www.arl.org/index.php Acesso em: 19 julho 2018.

uma biblioteca digital como uma biblioteca. Assim, ela deverá incluir serviços de referências com serviços de alerta, manter o banco de dados com perfil de pesquisa dos utilizadores, auxiliando-os com as ferramentas de pesquisa, acesso e assistência. A informação nela armazenada precisa de ser de alta qualidade bem como passar pelo processo de seleção, indexação, catalogação e classificação. A biblioteca visa, também, procurar a exatidão e integridade das fontes de informação nela disponibilizada e ter em conta a preocupação com a correta identificação.

A UNESCO aprova o Manifesto da IFLA para Bibliotecas Digitais (2011), apresentando princípios para ajudar as bibliotecas na realização de atividades de digitalização sustentáveis e interoperáveis para superar o fosso digital. Um fator crucial para alcançar os Objetivos de Desenvolvimento do Milénio das Nações Unidas é atenuar a exclusão digital. O acesso a recursos de informação e aos meios de comunicação apoia a saúde e a educação tanto quanto a cultura e desenvolvimento económico.

Segundo o Manifesto suprarreferido, uma biblioteca digital é uma coleção online de objetos digitais de qualidade garantida, que são criados ou recebidos e geridos de acordo com princípios internacionalmente aceites para o desenvolvimento de coleções e acessíveis de uma forma coerente e sustentável, apoiado por serviços necessários para permitir aos utilizadores recuperar e explorar os recursos. É ainda uma parte integrante dos serviços de uma biblioteca, aplicando novas tecnologias para fornecer acesso a coleções digitais. As coleções digitais são criadas, geridas e disponibilizadas de tal forma a serem facilmente e economicamente disponíveis para uso da comunidade onde está inserida.

Para cumprir a missão de acesso direto a recursos de informação digital, de forma estruturada e autorizada devem cumprir-se os seguintes objetivos:

1. Apoiar a digitalização, acesso e preservação do património cultural e científico;
2. Proporcionar acesso a todos os utilizadores aos recursos de informação recolhidos pelas bibliotecas, respeitando os direitos de propriedade intelectual;
3. Criar sistemas interoperáveis de biblioteca digital para promover padrões de livre acesso;
4. Apoiar o papel essencial das bibliotecas e serviços de informação na promoção de normas comuns e as melhores práticas;

5. Criar a consciência da necessidade urgente de garantir a acessibilidade permanente do material digital;
6. Ligar as bibliotecas digitais de pesquisa de alta velocidade e redes de desenvolvimento;
7. Aproveitar-se da maior convergência de meios de comunicação e papéis institucionais para criar e disseminar conteúdo digital.

> Digital libraries should represent the evolution and extension of traditional physical libraries in terms of the available information, both in formats and in extension, (...) Digital Libraries basically store materials in electronic format to be then accessed and possibly manipulated by different kinds of users, very often connected via the Internet (Calvanese, 2000).

Para (Giordano, 2016) *apud* SAYÃO, 2009, p.9) a Biblioteca digital é um ambiente que integra coleções, serviços e pessoas na sustentação do ciclo complexo de criação, disseminação, uso e preservação de dados, informação e conhecimento.

1.3 As Hemerotecas Digitais

Analisando o significado etimológico, "hemeroteca é um termo de origem grega, onde *heméra* significa "dia" e *thèke*, significa "depósito ou coleção" (Medeiros, Melo, & Nascimento, 1990). É de frisar que uma hemeroteca, para além da organização adequada para facilitar a realização de consultas, exige que a biblioteca disponha de espaço físico para o armazenamento destes documentos. Hemeroteca refere-se a um acervo de jornais e revistas, de modo a que apresente uma determinada organização técnica que facilite o processo de pesquisa e recuperação da informação. No seu processo organizacional, fez-se uso de técnicas documentais como, por exemplo, a indexação de assunto, tanto de modo genérico, quanto específico (Medeiros, Melo & Nascimento, 1990).

Na conceção de (Oliveira, 2005) a hemeroteca é designada como "Seção das bibliotecas em que se colecionam jornais e revistas". Este autor, ao discorrer sobre a etimologia do termo "hemeroteca", destaca o prefixo *hemi* como correspondente a "meio", ou seja, "pela metade". Refere-se a qualquer coleção ou conjunto organizado de periódicos, podendo ser uma secção de uma biblioteca apenas reservada à conservação deste tipo de documentos.

Há séculos que a preocupação com a estruturação, representação e organização de conteúdos informacionais acompanhou o desenvolvimento cultural na sociedade. No entanto, é na contemporaneidade, com a rotura das barreiras de tempo e espaço desmistificadas pelas TIC

que essa preocupação adquiriu maior importância, principalmente, depois do aparecimento da Internet e do ambiente *WWW*. Estas duas tecnologias, cujas interfaces computacionais têm vindo a interferir na forma como produzimos e acedemos à informação, resulta de uma constante modificação de um sistema híbrido, juntando o Homem à máquina, aumentando exponencialmente o fluxo informacional, favorecendo o efetivo acesso à informação e consecutivamente ao conhecimento, tornando-os assim, bens indispensáveis para o desenvolvimento global da humanidade.

Posto isto, em contexto documental, uma hemeroteca constitui-se como uma fonte de informação alternativa, seja impressa ou digital. Entende-se que, hemeroteca se refere a um acervo de jornais e/ou revistas, de modo a que apresente de uma determinada organização técnica que facilite o processo de pesquisa e recuperação da informação. Em termos organizacionais (Medeiros et al., 2008) faz-se uso de técnicas documentais como, por exemplo, a indexação de assunto, tanto de um modo genérico, como específico. Geralmente, a formato de uma hemeroteca impressa é feito por assunto ou título e o seu armazenamento é realizado em pastas suspensas ou em caixas de arquivo, podendo também passar por uma encadernação (Medeiros et al., *apud* Oliveira, 2005). Este género de hemeroteca exige bastante espaço por partes das bibliotecas ou centros de documentação, o que a pode tornar inviável para a instituição.

Uma hemeroteca não é apenas formada pela totalidade das edições de jornais e revistas, mas também é muitas vezes constituída com base em determinado assunto de interesse, dependendo do objetivo a ser atingido (Lampoglia, 2012). As formas de organização da informação são geralmente feitas por título ou por assunto, mas dependendo do sistema em que se encontra inserida, pode ser recuperada por data ou por palavra-chave. Para Carvalho (2006), nos últimos anos, a importância dos jornais para os estudos académicos tornou-se ainda maior graças à digitalização de um grande número de títulos, dos jornais pequenos e locais aos grandes e com grande circulação internacional, dos não correntes àqueles que ainda se encontram nas ruas. A partir dos anos 2000, os historiadores estavam diante de um universo quase inesgotável de reportagens, notícias, editais, anúncios, notas, colunas sociais e crónicas. Ao mesmo tempo, bibliotecas, museus, arquivos e universidades viram-se deparados com a necessidade de estabelecer critérios para o tratamento desse material.

> 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 económicas e políticas (Bauer, 1970, p. 85).

Para Lampoglia (2012, p.126), a hemeroteca permite a socialização dos dizeres do passado, instigando aos sujeitos repensar a história, procurar o contexto em que os escritos foram produzidos, comparar o passado com o presente e observar o que mudou e o que continua nos discursos da/sobre a história.

1.4 As questões de preservação e disseminação das coleções

À medida que os suportes de informação se diversificam, a disseminação da informação torna-se complexa. A conceção de uma biblioteca digital deve ser vista como uma ferramenta para proporcionar o acesso à informação constituída em ambiente digital e também deve, contudo, constituir-se como um instrumento para a democratização do acesso ao conhecimento (Rosetto, 2006). As bibliotecas digitais servem como ferramenta de acesso e disseminação por meio dos utilizadores. Para Cavalcante & Cavalcante (2010) a informação é pensada como um fator de produção dispensável ao processo de construção do conhecimento e de desenvolvimento das pessoas e das organizações. Pode ser classificada, analisada, estudada e processada de qualquer outra forma. Disseminar a informação pode ser entendida como "propagação, ou seja, a informação sendo difundida por vários meios e suportes abrangendo um determinado utilizador com base no esquema tradicional de emissor, canal, mensagem, recetor" (Cavalcante & Cavalcante, 2010). A preservação é um dos grandes desafios do séc. XXI. Durante os últimos anos do séc. XX, apenas as bibliotecas, os aquivos e os centros e institutos de pesquisa e organismos governamentais criavam conteúdo digital relevante (Arellano, 2004). Todas as instituições que estejam voltadas para a disseminação da informação, como é o caso das bibliotecas públicas, encontram-se num momento de questão sobre quais as práticas estabelecidas, tal como se questionam sobre as novas possibilidades das TIC. A distância e o tempo entre a fonte de informação e o seu destinatário deixaram de ter qualquer importância; os utilizadores não precisam de se deslocar uma vez que os dados viajam (Santos, 2004). Atualmente disseminar informação através do uso intensivo das novas tecnologias equivale a pensar na transmissão de bits. O formato digital da informação representa ao mesmo tempo uma nova maneira de, conforme Santos (2004), apresentar conteúdos antigos, como também, a partir desta combinação inteiramente nova de fontes, criar a possibilidade para que um novo conteúdo venha a ser criado, como a interatividade na consulta de conteúdos. Cada biblioteca deve refletir as suas

próprias necessidades de preservação das coleções. Os requisitos de preservação de uma biblioteca pública são diferentes dos outros tipos de bibliotecas. Contudo, todas são obrigadas a manter e a garantir a acessibilidade das suas coleções, quer por alguns anos, quer indefinidamente. De um ponto de vista económico, as bibliotecas não podem suportar um desgaste prematuro dos seus fundos.

> Replacing library material, even when possible, is expensive. Preservation makes good economic sense. Preserving current collections is the best way to serve future users (Adcock, 1998).

As bibliotecas e outras instituições assumiram a responsabilidade pela preservação da informação face a aspetos técnicos, legais e organizacionais em resposta às novas exigências da preservação digital. Exemplo disso é o RLG[3] (*Research Libraries Group*), que financiou um estudo das Metodologias e das necessidades para a preservação digital das instituições dos membros associados: *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information* (1996), que se baseia na ideia da preservação digital, chamando a atenção para a necessidade de um número suficiente de organizações capazes de armazenar, migrar e fornecer acesso aos seus arquivos digitais como elementos essenciais de um sistema. A preservação digital compreende os mecanismos que permitem o armazenamento em repositórios de dados digitais que garantiriam a perenidade dos seus conteúdos; as condições básicas à preservação digital seriam, então, a adoção de métodos e tecnologias que integrariam a preservação física, lógica e intelectual dos objetos digitais (Arellano, 2004). Uma maneira simples de começar um processo de preservação de dados é pelo uso do sistema *handle*, que mantém um endereço no ar, facilitando que o documento seja encontrado e a sua identificação seja feita. O sistema fornece serviços de resolução eficientes e segura para identificadores exclusivos e persistentes de objetos digitais, e é um componente do *Digital Object Architecture, do Corporation for National Research Initiatives*[4]. Fornece uma forma de gerir informações digitais num ambiente de rede. Um objeto digital tem uma máquina e uma plataforma de estrutura independente, que permite que ele seja identificado, acedido e protegido, conforme for apropriado. Um objeto digital pode incorporar não apenas os elementos informativos, ou seja, uma versão digitalizada de um papel, filme ou gravação de sim, mas também o identificador único do objeto digital e

[3] O RLG desenvolveu o mecanismo de pesquisa entre bibliotecas da Universidade de Nova Iorque nos EUA. Disponível em: http://www.rlg.org/ Acedido em: 3 agosto 2018.

[4] Corporação para Iniciativas Nacionais de Pesquisa é uma organização sem fins lucrativos formada em 1986 para empreender, fomentar e promover pesquisas de interesse público. As atividades concentram-se no desenvolvimento estratégico de tecnologias de informação baseadas em rede. Disponível em: https://www.cnri.reston.va.us/ Acesso em: 21 julho 2018.

outros metadados sobre o objeto digital (Araujo & Boeres, 2012). Os metadados podem incluir, se for o caso, restrições de acesso aos objetos digitais, avisos de propriedade e os identificadores para acordos de licenciamento.

A estabilidade química e física do material da Biblioteca também está dependente da qualidade e do processamento das matérias-primas utilizadas no seu fabrico, assim como da conceção e da montagem. Através dos séculos, as pressões colocadas pela produção em massa reduziram a qualidade do material recebido nas Bibliotecas. Muito do papel fabricado após 1850 é altamente ácido, estando a tornar-se mais quebradiço, o que fará com que se destrua mais facilmente com o tempo. As técnicas de encadernação foram simplificadas com a mecanização e muitos cadernos de texto são agora são muitas vezes colados com adesivo.

Deve, de igual modo, ter-se em conta os meios de suporte mais modernos como, os microfilmes, as fotografias e meios audiovisuais e os documentos em formato digital, possuem todos problemas de preservação inerentes e precisam de ser armazenados e utilizados com cuidado, caso se pretenda que não sejam prematuramente destruídos. De uma forma geral, é difícil aceitar o fato de uma grande quantidade de material de biblioteca estar a chegar ao fim da sua vida natural e de os poucos anos que ainda lhe restam só poderem ser prolongados através de um manuseamento e de um armazenamento cuidadosos.

Muitos países criaram programas de digitalização nacional e outros acordaram fazê-lo na Cimeira Mundial sobre a Sociedade da Informação (*World Summit on the Information Society*). Também conhecida como Cúpula Mundial ou Conferência Global das Nações Unidas, é caracterizada pelas suas reuniões de alto nível, que incluem Chefes de Estado e são intergovernamentais. A ideia de realizar esta cimeira surgiu em 1998, por iniciativa da Conferência Plenipotenciária da União Internacional das Telecomunicações[5]- UIT, a qual admitiu que o fosso digital (*digital divide*) existente entre indivíduos com e sem acesso à informação estava a disparar, em oposição à evolução das Tecnologias de Informação e Comunicação (TIC), as quais passaram a desempenhar um papel cada vez mais relevante nos âmbitos político, social, cultural e económico. Ocorreu a 21 de dezembro de 2011 a aprovação para a realização da CMSI, através da Assembleia Geral das Nações Unidas, a qual recomendou que os preparativos para a cimeira fossem realizados através de um Comité

[5] A União Internacional de Telecomunicações-UIT é uma organização internacional com a função de padronizar e regular as ondas de rádio e telecomunicações internacionais. A UIT está dividida em três Sectores: Radiocomunicações, Normalização das Telecomunicações e Desenvolvimento das telecomunicações. Disponível em: https://www.anacom.pt/render.jsp?contentId=65260 Acesso em: 21 julho 2018.

Preparatório Intergovernamental em aberto, o qual deveria: definir a agenda, decidir sobre as modalidades de participação dos outros setores interessados (Horizonte, 2012).

É inegável a importância da digitalização de acervos antigos para facilitar o acesso a informações que antes só poderiam ser consultadas em espaços físicos e únicos e nem sempre de acesso fácil aos utilizadores. A digitalização colabora para a preservação dos originais à medida que os utilizadores podem recorrer à cópia em versão digital e não à versão em papel. Em muitos casos, no que está relacionado com a coleção de jornais e revistas, os exemplares guardados são cópias únicas, representando um verdadeiro tesouro de memória de uma determinada época (Giordano, 2016).

> Quick, reliable, comprehensive access to information in newspapers has long been a recognized need. Libraries, historical societies and news organizations have attempted to meet that publicly expressed need in a variety of ways over the years, from bound ledgers, loose-leaf notebooks and card files to keypunch sorters, mainframe computers and personal computers. The major change in the last decade has been the advent of online full-text newspaper databases. Although most of these formats are still in use today, the trend is toward computer-assisted indexing (Semonche, 2003).

A reorganização eletrónica de um *clipping archive* deve fornecer as mesmas funções que um *analogue archive*. Os artigos devem ser selecionados, recortados, reorganizados, arquivados e disponibilizados aos utilizadores como revisões. Com base no estudo do *LAURIN PROJECT* os jornais a digitalizar são selecionados segundo os dados essenciais, como o nome do jornal, a edição, o número da página e a data de publicação. Os artigos podem ser recortados eletronicamente da página que se está a digitalizar e são reconstruídos numa página *standard*.

O esboço geral dos desafios da preservação digital está bem estabelecido; os materiais digitais são especialmente vulneráveis à perda e destruição dada a natureza dos próprios materiais, desastres naturais e provocados pelo Homem, o ambiente onde os materiais são mantidos, o modo como os materiais são manuseados. Segundo as Diretrizes da IFLA para a conservação e o manuseamento de documentos da Biblioteca (IFLA, 2014):

> Traditional library collections contain a wide range of organic materials, including paper, cloth, animal skins, and adhesives. Such organic substances undergo a continual and inevitable natural ageing process. While measures can be taken to slow this deterioration by careful handling and providing a sympathetic environment, it is impossible to halt it altogether (IFLA, 2014).

Oferecer coleções de jornais em formato digital e disponível na Web reduz a necessidade de deslocação dos utilizadores, aumenta o interesse e permite a pesquisa no momento mais

adequado para o utilizador, podendo este escolher quando e de onde aceder ao acervo digitalizado.

(Zogla, 2014) demonstrated two main benefits of newspaper digitization:

- Anyone could access a newspaper archive from their home computer.
- By combining holdings of several libraries, full collections of some newspapers could be created in a digital form that no library on its own held in a physical collection.

CAP 2 REVISÃO DE PROJETOS DE HEMEROTECAS DIGITAIS

2.1 Projetos

Foram escolhidos os projetos NDNP, Europeana e MoA para se formar o modelo de construção de uma hemeroteca de jornais. O que mais carateriza estes projetos é o seu caráter heterogéneo no que concerne às suas coleções. Estes projetos são conhecidos por produzirem conteúdo bibliográfico que pode ser consultado através do utilizador nas suas respetivas páginas online. Foi através desse conteúdo que foram extraídos os processos de digitalização, processamento e publicação de uma coleção de jornais, que mais tarde foram interpretados e adaptados aos objetivos deste projeto. Os três projetos têm um carater versátil, flexibilidade, coerência e simplicidade pelo que, pode ser adaptado às funcionalidades do alinhamento com os objetivos criados por cada instituição.

2.1.1 Qual é a importância de ter um modelo de construção de uma Hemeroteca Digital?

A digitalização de documentos ajuda a preservar o acervo físico, facilita o acesso à informação e ajuda na partilha de conhecimento. Um documento histórico, como o jornal, só poderá ser consultado no local onde está armazenado, isto implica deslocações e manuseamento que poderá ser prejudicial para o material fragilizado. Assim, se uma instituição se consciencializar da importância que um guia de boas práticas, que conduza à criação de uma hemeroteca de jornais tem, irá reunir aspetos positivos para quem quer consultar, a uma longa distância e através de um clique e para a própria instituição. Melhorar a segurança do acervo é outro dos grandes aspetos a considerar. Acidentes acontecem e podemos assistir a perdas irreparáveis, como é o caso da possível ocorrência de um incêndio. Se tivermos o acervo digitalizado, garantimos a preservação da memória cultural e social que o jornal tanto representa.

.2.2 Revisão de Projetos

2.2.1 *National Digital Newspaper Program*

A digitalização de jornais é uma prática muito recente, existem instituições que têm a função de difundir e salvaguardar a memória que ainda estão a aprimorar a sua atividade. A digitalização é um processo que leva algum tempo, exige investimento financeiro, profissionais capazes de se entregar a estas tarefas e utilizadores que apoiem os suportes.

Na década de 1990 quando os primeiros jornais começaram a ser digitalizados nos Estados Unidos da América, não faltavam opções tecnológicas no mercado. Umas eram de ótima qualidade, enquanto outras apresentavam resultados pouco satisfatórios (Carvalho, n.d.). Os jornais digitalizados eram condicionados a sistemas de marca registada, uma vez que algumas opções tecnológicas pertenciam às empresas que as comercializavam, querendo isto dizer que não podiam ser transferidas para o comprador. Assim, todo este processo implicaria gastos constantes para as instituições: bibliotecas, arquivos, museus e universidades estavam presas a um modelo pouco flexível e bastante dispendioso (University of California, 2011).

Foi então que em 2005 surgiu o Programa Nacional de Jornais Digitais (NDNP), uma parceria criada entre o *National Endowment for Humanities* (NEH) e a Biblioteca do Congresso, construído a longo prazo para fornecer acesso permanente a um recurso digital nacional de informação bibliográfica de jornais e jornais históricos, selecionados e digitalizados por instituições financiadas pelo NEH.

> The National Digital Newspaper Program supports a consistent technical specification for digital newspaper reproductions and associated metadata to maintain parity of services for materials from a variety of institutions and collections and to support the "best practices" of today's understanding of digital preservation needs (Digital & Program, 2009).

Este projeto que visava a colaboração de todos os estados dos Estados Unidos da América tinha duas grandes metas: criar uma plataforma digital que reunisse todos os jornais de todos os estados americanos: *Chronicling America* e criar um portfólio de melhores práticas para a digitalização de jornais.

> The NDNP has two distinct but related goals. The first is to create a free online resource of searchable newspapers from every state for the period 1836-1922, available through the Chronicling America website (http://chroniclingamerica.loc.gov/). Newspapers hosted here are true digital facsimiles of the original newspapers and are full-text searchable.
>
> The second goal of the NDNP is to create a set of best practices for newspaper digitization. Libraries have always recognized the importance of using open standards to facilitate collaboration and data sharing. (Digitizing California's Newspapers: A Guide and Best-Practices for Institutions Around the Golden State, 2011)

Para a compreensão do NDNP temos como exemplo dois dos primeiros projetos que receberam um financiamento de apoio à digitalização de jornais, o *Utah Digital Newspapers* (UDN) em 2005 e o T*he Center for Bibliographical Studies and Research* (CBSR) com a Universidade da Califórnia que lançaram o *California Digital Newspaper Collection* (CDNC) em 2006.

Com o intuito de digitalizar e disponibilizar na internet, de forma gratuita, os jornais históricos do estado do Utah, em parceria com a Universidade do Utah, a Universidade *Brigham Young* e

a Universidade do Estado de Utah. Algumas das diretrizes usadas através do NDNP pelo UDN foram, a utilização de mapas para apontar a origem geográfica dos jornais, a indexação de todo o conteúdo, a colaboração de diversos utilizadores para o projeto e o uso essencial dos microfilmes para proceder à digitalização. A primeira etapa é criar imagens digitais de 400 dpi (ponto de polegada), em tons de cinza e em formato TIFF (*tagged image format*). O software corta as imagens – de forma a eliminar as arestas no quadro que não fazem parte da página – faz certas deslocações de forma a orientar a página diretamente para cima ou para baixo e, por último, faz uma edição da imagem de forma a "limpar" os borrões do scanner. Cada página digitalizada passa por um processo de identificação e classificação de cada parte:

> 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 (Herbert & Estlund, 2008).

O OCR entra em ação, uma vez que este é um processo automatizado que executa o reconhecimento de caracteres em cada artigo de uma página digitalizada, criando um arquivo do texto do arquivo. John Herbert e Karen Estlund (2008) afirmam ainda - *After generating the "raw" text, another automated process filters it through English dictionaries, a Utah place-names dictionary, and an extensive surnames list*. Quanto aos títulos e subtítulos, estes são transcritos à mão de forma a garantir a sua precisão. A imagem de cada artigo é armazenada em arquivos separados, bem como a imagem da página inteira, as imagens são posteriormente gravadas em arquivos PDF.

O CDNC seguiu igualmente orientações técnicas e metodológicas inspiradas nos manuais NDNP como, avaliar a qualidade e quantidade do conteúdo disponível, determinar quais ou títulos mais valiosos (tendo em conta a limitação dos fundos disponíveis), privilegiar jornais publicados entre 1923 (por causa dos direitos de autor), procurar digitalizar jornais microfilmados em detrimento do material impresso - a digitalização de microfilmes é mais eficiente e mais em conta. É importante, produzir cópias de segurança e de seguida elaborar inventários, metadados e assegurar que sejam usados formatos digitais compatíveis.

2.2.2 Europeana

> A Europeana nasceu da necessidade de unir diversos documentos espalhados pelas bibliotecas, museus, arquivos e outros centros culturais espalhados pela Europa, numa tentativa de disponibilizar num ambiente digital o acesso a essa grande massa informacional (Coneglian & Santarem Segundo, 2016).

O Projeto da Europeana para a digitalização de jornais oferece possibilidades de pesquisa e navegação pelos jornais de toda a Europa em formato acesso aberto para todos os interessados,

pesquisadores, estudantes e professores. Este projeto financiado pela Comissão Europeia conta com cerca de 18 milhões de páginas digitalizadas de jornais em texto completo. Durante o projeto, instituições e bibliotecas nacionais forneceram conteúdo para a construção para se tornar o maior fornecedor de coleções de jornais europeus de forma a criar uma identidade europeia, em contexto político e histórico. Nesse sentido, o projeto fornece uma pesquisa complexa do conteúdo completo de texto, que inclui o uso de diversas ferramentas e tecnologias para a pesquisa e recuperação da informação (Milnovic, Trtovac & Sofronijevic, 2014).

Refletindo sobre a criação do conjunto de dados, foram necessárias várias centenas de profissionais para selecionar, pré-processar, corrigir manualmente, verificar, ingerir e categorizar as imagens da página e o conteúdo de verdade do terreno. Além disso, exigiu especificações rigorosas e detalhadas e um controlo rigoroso de qualidade. Um aspeto significativo que tornou este esforço único foi a colaboração de um grande número de bibliotecas nacionais e de outras importantes bibliotecas europeias para tornar este conjunto de dados representativo e gratuito (Clausner, Papadopoulos, Pletschacher, & Antonacopoulos, 2015).

Para Rossitza Atanassova (2015) a atual versão da interface do projeto incorpora funcionalidades básicas de pesquisa e navegação associadas aos sites de jornais digitalizados. Os utilizadores podem pesquisar através de palavras-chave, por texto completo ou pelos títulos dos jornais. A pesquisa pode ser melhorada através de "content provider, language and a set of years of publication, or browse by newspaper title, date and contributing country" (Atanassova, 2015). Desde que a interface forneça o acesso a conteúdo multinacional, a agregação e o *design* do navegador levam em consideração quaisquer restrições impostas pelas bibliotecas contribuintes devido às leis nacionais de direitos de autor e aos modelos de digitalização de jornais das bibliotecas.

Falando no desempenho daqueles que participaram (investigadores na sua maioria) nos testes de usabilidade e no *feedback* das entrevistas, foi demonstrado que o método preferido de usar o arquivo digital de jornais históricos é através de opções de pesquisa controladas e não através de uma navegação simples; embora este grupo específico de utilizadores gostasse que a Europeana disponibilizasse opções de pesquisa mais avançadas, ajudando a filtrar e a manipular os resultados de pesquisa. Essas funcionalidades seriam consistentes com as opções de pesquisa avançada - pesquisa Booleana, facetas por tipo de artigo ou camadas de filtros - implementadas noutro tipo de interfaces como no *Chronicling America*, no Arquivo de Jornais Britânico ou no Trove. Outra das opções apresentadas pelos utilizadores mencionava a pesquisa por assunto e ainda por período histórico, o que aponta para um interesse muito específico e não geral no site.

O principal grupo de utilizadores da Europeana são os investigadores, como foi confirmado pela primeira ronda de testes de usabilidade, há realmente uma grande procura de jornais históricos digitalizados e não é surpresa para a comunidade *Europeana Newspapers.*

> The researchers interviewed turn to newspapers to study a range of subjects and topics: 19th century popular culture and humour, history, literature, evolution of language, public discourse, reference cultures and professional careers. For them such an aggregation of millions of pages of European newspapers offers exciting new opportunities for "transnational comparative research" and computational analysis of the data (Atanassova, 2015).

Quando os jornais são digitalizados, a versão eletrónica resultante é muitas vezes uma simples imagem do jornal. Nem sempre é possível pesquisar efetivamente imagens, artigos ou termos individuais no texto. O projeto da *Europeana Newspapers* deteta e marca milhões de artigos únicos com metadados relacionados e entidades nomeadas (*Named-Entity Recognition* – NER). O principal objetivo do projeto é pesquisar as coleções existentes de jornais digitalizados e de materiais para agregar metadados sobre eles. Para além do mais, visa o aperfeiçoamento das coleções já existentes de materiais digitalizados através da tradução de materiais de imagem em texto completo por meio de OCR – reconhecimento ótico de carateres e OLR – reconhecimento ótico de layout.

Hans-Jörg Lieder, da Biblioteca Estadual de Berlim e coordenador do projeto, referiu que*:*

> Technologies used during assessments of quality and refinement of newspaper articles - optical character recognition (OCR), optical layout recognition (OLR), named entity recognition (NER) and class recognition (Trtovac, Republic, & Jerkov, 2014).

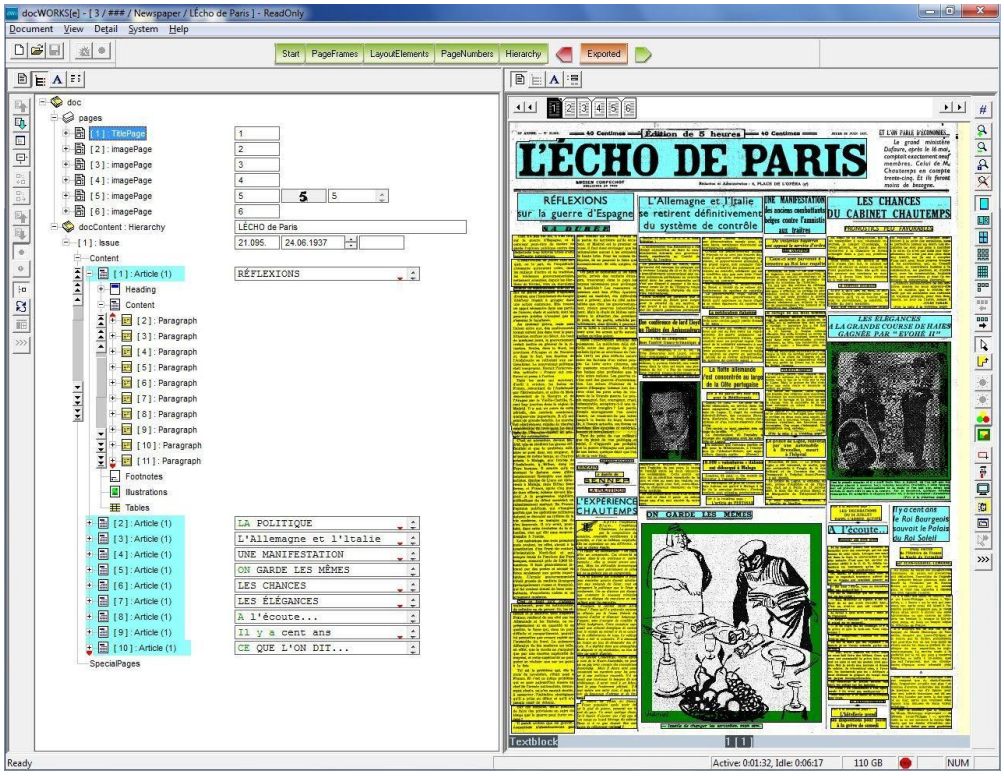


Figura 1 A newspaper separated into articles through techniques such as Optical Layout Recognition (OLR)
FONTE: EUROPEANA

Contudo, o autor enfatizou a importância do projeto em termos de metadados de artigos de jornais, para além da pesquisa de texto completo, é possível pesquisar através de metadados cujos textos descrevem as fotos com as quais os artigos dos jornais são ilustrados. Lieder falou também sobre os desafios e as dificuldades encontradas durante o refinamento de textos em jornais antigos. Em primeiro lugar, os jornais antigos são impressos com tinta de baixa qualidade, o que acaba por dificultar o reconhecimento ótico de carateres. Para os utilizadores que usem a Europeana, esta inclui a capacidade de realizar pesquisas através de palavras-chave e pesquisa de frases, navegação através de imagens e informação sobre elas (Trtovac, Republic, & Jerkov, 2014).

A abordagem adotada segundo o *The IMPACT Dataset of Historical Document Images* (Papadopoulos, Pletschacher, Clausner, & Antonacopoulos, 2013) para construir o conjunto de dados deu ênfase a um conjunto de matéria: realista – refletindo as atuais propriedades da biblioteca no que diz respeito à representatividade e frequência de documentos; abrangente – incluindo metadados e grandes detalhes sobre o que é verdadeiro; estruturado de forma flexível - apoiando as partes interessadas para pesquisar, navegar, agrupar, etc. e permitindo outros sistemas técnicos como os sistemas de fluxo de trabalho e avaliação das ferramentas.

> There is a significant need to explicitly record information of diverse nature used or produced by methods within digitization pipelines – both for making the substitution of alternative methods possible and for evaluating intermediate stages (Pletschacher & Antonacopoulos, 2010).

O formato de um *framework* que atenda a essa necessidade deve ser capaz de fornecer informação detalhada e precisa sobre os resultados de cada utilizador na etapa de processamento, bem como informações cumulativas sobre os resultados dos processos que ocorreram em determinado ponto. Existem vários formatos que foram desenvolvidos predominantemente para a gravação de resultados de análise e reconhecimento. Tais formatos são adequados para armazenar os resultados finais da análise de documentos e alimentar sistemas de publicação: ALTO ou hOCR. *Technical Metadata for Layout and Text Objects* (ALTO) é o método de análise e objeto de texto que segue um esquema XML que detalha metadados técnicos para descrever o layout e o conteúdo de recursos, codificando o conteúdo textual de uma página digitalizada em grande detalhe, incluindo estilos e *layouts*.

> Over the history of OCR systems, a significant number of formats have been proposed for representing the output of OCR systems. We can distinguish three major classes of OCR output formats: logical formats, suitable for direct use of OCR results by end users (RTF, HTML, LaTeX, and Microsoft Word), OCR engine-specific formats, and benchmarking formats proposed for benchmarking various aspects of OCR systems (Breuel & Kaiserslautern, 2007).

A análise de *layout* é a primeira etapa, e uma das mais importantes, na análise de imagem de documento, onde, após o aperfeiçoamento da imagem, é obtida uma representação descritiva da estrutura da página. A análise de *layout* é fundamental para a maioria dos sistemas e aplicações de análise de imagens de documentos, compreende a segmentação de página – identificação de regiões de interesse, classificação de região – identificação do tipo de conteúdo de cada região e outros processos como rotulagem lógica – de regiões em termos da sua função. Durante vários anos a ICDAR[6], dedicou-se consideravelmente para desenvolver diversos métodos de análise de *layout*, em particular, a segmentação de página. A maioria dos métodos foi destinada a aplicações específicas e, consequentemente, foram baseadas nem suposições específicas nas suas classes de documentos de destino, como por exemplo em blocos de texto. Cada método foi avaliado em conjuntos de dados específicos de aplicações com foco estreito e que geralmente não refletem a ocorrência de documentos no mundo real. As diferentes abordagens de anos anteriores da ICDAR, concentrara-se no cálculo de várias métricas de erro

[6] Conferência Internacional sobre Análise e Reconhecimento de Documentos, é o maior e mais importante encontro internacional de pesquisadores, cientistas e profissionais da comunidade de análise de documentos. Disponível em: https://ieeexplore.ieee.org/xpl/conhome.jsp?punumber=1000219. Consultado em: 21 outubro 2018.

para quantificar o desempenho dos métodos de segmentação de página, principalmente para *benchemarking* ou avaliação comparativa. Estas primeiras abordagens consideraram o texto reconhecido dentro de cada região e o número corresponde de operações de edição necessárias para corrigir erros. No entanto, esta métrica não pode fornecer indicações precisas sobre o desempenho de segmentação de página, uma vez que, muitos dos erros no texto são provocados pelo processo de OCR (Antonacopoulos & Bridson, 2007) . Existem regiões homogéneas que são identificadas, tais como a segmentação da página que são rotuladas de acordo com o tipo de conteúdo. A exatidão do resultado da segmentação de página e da classificação da região é fundamental, uma vez que a representação resultante forma uma base para os processos seguintes.

> Layout Analysis is one of the most well-researched fields in Document Image Analysis, yet new methods continue to be reported in the literature, indicating that the problem is far from being solved (Antonacopoulos, Clausner, Papadopoulos, & Pletschacher, 2013).

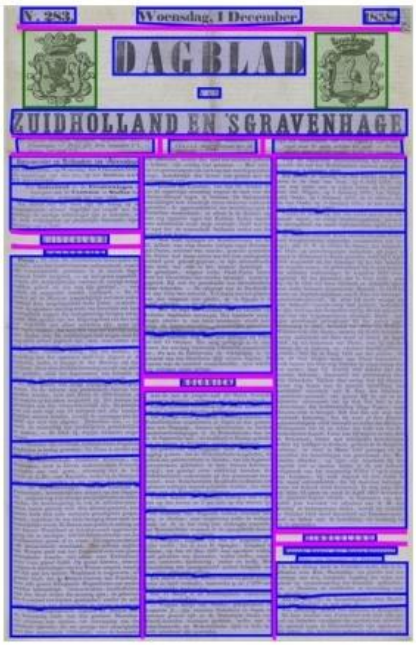
Um conjunto abrangente de dados de imagens de documentos históricos foi criado como parte do projeto IMPACT[7], estando disponível através do Centro de Competência em Digitalização. Este conjunto de dados foi coletado para refletir não apenas as condições e os artefactos de documentação histórica que impliquem a base de dados, como também as necessidades e prioridades das bibliotecas, em termos de quais os tipos de documentos dominam os seus planos de digitalização. O projeto é constituído por 15 diferentes detentores de conteúdo, incluindo a maioria das bibliotecas nacionais e principais da Europa. Para o propósito da competição da Europeana para a Análise de *Layout* de Jornais Históricos, foram selecionadas um conjunto de imagens de dados do IMPACT como amostra representativa de diferentes idades garantindo a presença de diversas questões que afetassem a análise do layout (Antonacopoulos et al., 2013).

> Such issues include dense printing (minimal spacing), irregular spacing, varying text column widths, presence of separators, interspersed graphics/adverts, presence of black borders, text printed in different orientations (horizontal and vertical) and different number of columns (from 2 to 6).

Um intervalo de metadados é registado para cada região diferente. O formato oferece meios sofisticados para expressar a ordem de leitura e relações mais complexas entre as regiões.

Como podemos observar na seguinte imagem (figura 2):

[7] IMPACT: é um projeto financiado pela Comissão Europeia. O seu objetivo é melhorar significativamente o acesso ao texto histórico e eliminar as barreiras que impedem a digitalização em massa do património cultural europeu. Disponível em: http://www.impact-project.eu/. Consultado em: 2 agosto 2018.



http://www.primaresearch.org/www/assets/papers/ICDAR2013_Antonacopoulos_HNLA2013.pdf. *Consultado em: 2 agosto 2018.*

Após todo este processo de Análise de *Layout* de Jornais Históricos, o objetivo foi avaliar os métodos apresentados num conjunto de dados de jornais históricos representativos, em termos de diferentes *layouts* e prioridades de digitalização de bibliotecas. Na 12ª Conferência Internacional de Análise e Reconhecimento nos Documentos: Competição sobre Análise de *Layout* para Jornais Históricos (2013) foram relatados dois cenários em que é avaliada a capacidade dos métodos de segmentar regiões com precisão e classificada a região de extração do texto

Os jornais apresentam desafios específicos devido ao seu tamanho e consequentemente o tamanho das imagens, à baixa qualidade de impressão e ao *layout*. Com o objetivo de alcançar avanços significativos na análise de imagens para este género de documentação histórica, é importante estar ciente de todos os desafios e idiossincrasias apresentadas pelo material. Há uma necessidade significativa de criação de conjunto de dados representativos de imagens, que requer uma seleção cuidadosa, envolvendo custos consideráveis. As tecnologias aplicadas para o aperfeiçoamento ou *refinement technologies* aplicadas incluem, software de OCR, análise de *layout*, NER (*Named Entity Recognition)* e reconhecimento de página (Clausner et al., 2015).

Em termos de formato de imagem e renderização, o projeto da Europeana decidiu coletar versões que continham a melhor qualidade possível ou aquelas que estariam mais próximas do original e sempre que possível os resultados OCR também eram fornecidos. Foram recolhidos

metadados essenciais para permitir a indexação e a pesquisa no repositório, mantendo o esforço de conversão e mapeamento dos registos. Os detalhes obrigatórios eram, o título, o principal idioma, a fonte original e a data de publicação. De pretexto opcional estavam informações sobre o tipo de letra, o modelo do scanner, artefactos da imagem e comentários sobre a qualidade do documento. Com a recolha dos conteúdos, estes foram examinados pelos autores e inseridos no sistema de banco de dados do repositório, envolvendo a alocação de IDs exclusivos do projeto para cada imagem, para a conversão de um formato de imagem padrão comum com compactação sem perdas, gerir cópias de visualização, análise de caraterísticas de imagem e metadados de cabeçalhos de arquivos.

Para a seleção do conjunto final de dados foram motivadas duas questões (Clausner et al., 2015):

> 1. To narrow the initial selections further down so as to be in line with the available resources (budget);
> 2. To maintain the representativeness of the individual datasets as far as possible.

No que respeita à representatividade, a distribuição de idiomas, scripts, páginas de título, páginas do meio, *layouts* de caraterísticas e períodos de tempo foi mantida o mais próximo possível da seleção original. Todas as imagens de página no conjunto de dados são em 300dpi ou 400dpi e há uma ampla distribuição de páginas em tons de cinza bitonal e coloridas. Independentemente dos arquivos de origem da imagem original, todas as imagens no conjunto de dados são armazenadas como arquivos TIFF (sem perder compressão). O *Ground Truth* pode ser descrito como o resultado ideal do fluxo de trabalho de OCR. É crucial para avaliar a saída dos métodos de análise de documentos para o que seria considerado o resultado mais correto.

> The creation of ground truth is typically a manual or (at best) semi-automated task due to the fact that current OCR engines are still far from being perfect (especially for historical documents).

Todos os ficheiros dos dados verdadeiros foram criados no formato PAGE[8] (*Page Analysis and Ground Truth Elements*) estabelecido e recomentado pelo IMPACT. De forma a aumentar a produtividade, os provedores de serviços foram fornecidos com:

> Preliminary-processed OCR output files in PAGE (produced by the authors, using a PAGE exporter tool based on the ABBYY FineReader Engine 10[9]) which they could

[8] PAGE – uma estrutura de representação baseada em XML para análise de documentos e resultados de reconhecimento.

[9] ABBYY FineReader Engine 10 – reconhecimento de documentos, captura de dados e processamento de idiomas. Disponível em: https://www.abbyy.com Acedido em: 4 agosto 2018.

> either correct or, depending on the quality of the material, discard and create all ground truth manually (Clausner et al., 2015).

Usaram também uma versão personalizado do Aletheia[10], sistema de produção *ground truth* produzido pelos autores, e informações detalhadas sobre como interpretar e representar elementos de conteúdo específicos.

Após de se liberar o potencial do conjunto de dados exclusivos de imagens de jornais com a verdade básica correspondente, decidiu-se seguir as melhores práticas na coleta de dados e troná-lo acessível através de uma *web-based* (repositório online). Mostrou-se muito útil na perspetiva de estabelecimento de um ponto comum de referência entre os parceiros do projeto, que possuem Ids únicos, pesquisar e identificar documentos com caraterísticas específicas, construir subconjuntos para experiências de avaliação individual e permitir que outros sistemas técnicos acedam diretamente aos recursos(Clausner et al., 2015).

> The web presence of the dataset comprises five sections:
> Introduction – Overview of the content with statistics on the hosted material and its usage.
> Dataset – Entry point for browsing the dataset per contributing institution and specific subsets defined.
> Advanced Search – Entry point for searching by metadata, image characteristics, and attachments.
> Cart – A tool for managing and exporting selections of images and ground truth files.
> Login – Management of user details and password (Clausner et al., 2015).

Após o desempenho do fluxo de trabalho de OCR, que foi empregado para produção de larga escala a Europeana resumiu num documento os resultados da avaliação (Pletschacher, Clausner, & Antonacopoulos, 2015). Para a produção *ground truth* todos os verdadeiros dados firam pré-produzidos usando o *FineReader Engine10*. Foi então eu os provedores de serviço corrigiram manualmente erros de reconhecimento (layout da página e do texto) e a ferramenta *PAGE Validator* garantiu a verdade de precisão no controlo de qualidade. Quanto à produção de resultados OCR, foi produzida usando o fluxo de trabalho de produção da Europeana Newspapers que incluiu o método de *binarization* (binarização) de imagens através do NCSR Demokritos e o Abbyy FineReader Engine 11. Os resultados de reconhecimento foram obtidos nos formatos ALTO XML e FineReader XML, que posteriormente, foram convertidos para o formato PAGE XML. Para além disto, todas as imagens de documentos foram processadas com o Tesseract, o software de OCR de código aberto e de última geração para permitir a comparação de dois mecanismos de OCR diferentes. Quanto ao desempenho de

[10] ALETHEIA - é um sistema avançado para análise, reconhecimento e anotação precisos e com boa relação custo-benefício de documentos digitalizados. Disponível em: http://www.primaresearch.org/tools/Aletheia Acedido em: 4 agosto 2018.

reconhecimento de texto dos sistemas de OCR, este pôde ser essencialmente medido pela comparação de arquivos de texto simples. Para uma avaliação mais justa, foi preciso executar as seguintes etapas de processamento (Pletschacher et al., 2015):

1) Normalização de texto

De forma a preservar a informação, o texto de *groud truth* foi transcrito o mais próximo possível do documento original. Em relação ao mecanismo OCR, a avaliação foi feita de forma mais realista, uma vez que a relação dos conjuntos de carateres é limitada, especialmente aqueles relacionados com documentos históricos. A *ground truth* e o resultado de texto foram normalizados utilizando regras como, as recomendações do MUFI[11] (*Medieval Unicode Font Initiative*), os carateres de aparência semelhantes são mapeados como um só, as ligaduras são expandidas em carateres individuais e os carateres de um idioma específico, semelhantes com outros idiomas, não são substituídos (Pletschacher et al., 2015).

2) Exportação de texto

Como o texto Unicode real está incorporado na hierarquia de elementos dos arquivos XML de PÁGINA, foi necessário serializar todos os fluxos de texto. Para este fim, uma ferramenta exportadora foi usada para extrair apenas o conteúdo textual em arquivos de texto simples. Esse processo teve que levar em conta a ordem de leitura das regiões de texto, de modo a chegar a uma serialização válida do texto contido em *layouts* muito complexos.

3) Avaliação

A avaliação do desempenho foi realizada usando a ferramenta de avaliação de texto em dois diferentes métodos:

- O método de *Bag of Words;*

- O método de precisão da palavra.

Para comparação, um total de oito combinações diferentes de arquivos de entrada foram processadas:

- Resultados de OCR baseados em imagens bitonais e originais;

- Formato XML, ALTO XML e FineReader;

[11]Midieval Unicode Font Initiative é um grupo de trabalho sem fins lucrativos de designers de fontes. Disponível em: https://folk.uib.no/hnooh/mufi/ Acedido em: 16 agosto 2018.

- Texto original e texto normalizado.

Todas as execuções de avaliação de desempenho baseado em layout foram realizadas a partir da ferramenta PRImA. Vários fatores foram levados em conta, resultando em tabelas de resultados:

- Perfis de avaliação diferentes que correspondem aos cenários mencionados;

- Resultados de OCR baseados em imagens bitonais e originais;

- Formato XML, ALTO XML e FineReader.

Após apresentada uma visão geral e detalhada dos resultados de avaliação que foram obtidos pelo fluxo de trabalho de produção da Europeana, poderá concluir-se que são de boa qualidade e adequados para serem usados em vários cenários. Além disso, as decisões técnicas tomadas durante a configuração do fluxo de trabalho de produção podem ser confirmadas através de uma série de observações, como o desempenho do OCR, de linguagens e alguns problemas específicos de layout (Clausner et al., 2015).

A Associação de Investigação de Bibliotecas Europeias (LIBER) lidera as atividades de divulgação da Europeana, sendo um enorme esforço e investimento por parte da Associação com o objetivo de fornecer uma infraestrutura de informação para permitir que a pesquisa das instituições da LIBER seja de classe mundial, em particular no contexto de reformulação da biblioteca de pesquisa (Reilly, 2013). Esta biblioteca está a ser reformulada pelo ambiente de informações digitais em constate mudança, bem como pela mudança na forma como a pesquisa é realizada. O ambiente de pesquisa está relacionado com a digitalização das coleções, a recolha de dados digitais e dados de pesquisa.

> This is especially true for digitised newspaper collections, where to truly realise the value of newspapers and their unique attributes, further refinement must be carried out. Making digitised content available in this way can allow the researcher to place a research question in a wider context by exploring high quality images of the content in question and also by linking to related content (Reilly, 2013).

Ao explorar o estado das coleções dos jornais digitalizados da Europeana, um dos objetivos do projeto é identificar outras coleções de jornais digitais. Isto foi conseguido através de um levantamento das bibliotecas membros da LIBER[12]. A pesquisa concentrou-se no título do jornal e no intervalo de tempo, nos metadados usados, na distribuição de dados, nas capacidades e na qualidade da digitalização, incluindo a tecnologia utilizada. Para além de nos fornecer uma visão geral das coleções que poderiam beneficiar das melhores práticas desenvolvidas pelas

[12] LIBER – Liga das Bibliotecas Europeias.

bibliotecas da Europeana, a pesquisa descobriu alguns possíveis problemas e lacunas[13] a serem abordados como forma de tornar o património jornalístico da Europa mais acessível, através da Europeana (Reilly, 2013).

Tal como o serviço de agregação, o portal e os recursos de pesquisa em texto completo desenvolvidos pelo projeto da Europeana, devem ser usados para expor um novo tipo de conteúdo, os jornais. Desenvolvimentos recentes no OCR, feitos através do projeto IMPACT12 estão a ser aplicados ao conteúdo de jornais de bibliotecas nacionais por toda a Europa. Segundo Reilly (2013), 18 milhões de páginas de jornais serão refinadas e disponibilizadas através do portal Europeana. Ele fará uso de métodos de refinamento para OCR, segmentação de artigo, LOR e NER. O que foi desenvolvido através da Europeana está a ser agora aplicado às coleções dos jornais europeus (Reilly, 2013):

1. Agregação: Quatro tipos de coleções de jornais digitais exigentes podem ser identificados:

a) Apenas imagens com metadados estruturais;

b) Imagens com metadados estruturais e texto completo para pesquisa – OCR;

c) Imagens com metadados estruturais, OLR e OCR;

d) Imagens com metadados estruturais, OLR e OCR e enriquecimento semântico.

2. Padronização de metadados. Uma pluralidade de formatos de metadados existentes serão identificados e soluções serão fornecidas à comunidade;

3. Melhores recursos de exibição. Tornais os jornais fáceis de pesquisar e disponibilizá-los online.

2.2.3 Making Of America

O Projeto *Making Of America* traduz-se num esforço colaborativo com a Universidade de Michigan e a *Cornell University*, para preservar e tornar acessível, através da tecnologia digital, um conjunto significativo de fontes primárias relacionadas com a história social americana. Usando resultados "não processados", nem corrigidos após o reconhecimento ótico de carateres das imagens de página e codificação SGML[14] (*Standard Generalized Markup Language*) da informação textual resultante em conformidade com o TEI[15] (*Text Encoding Initiative*). Este

[13] Para mais informações consultar: http://www.europeana-newspapers.eu/public-materials/publications/

[14] SGML (Standard Generalized Markup Language) - é um padrão internacional abertamente documentado e livremente implementável para a marcação semântica de elementos textuais. Disponível em: http://www.loc.gov/preservation/digital/formats/fdd/fdd000465.shtml Acedido em: 8 agosto 2018.

[15] TEI (Text Encoding Initiative) - é um consórcio que desenvolve coletivamente e mantém um padrão para a representação de textos em formato digital. Disponível em: http://www.tei-c.org/ Acedido em: 9 agosto 2018.

projeto fornece acesso às imagens de página na Web sem ferramentas de visualização especiais, através de um sistema de entrega de páginas que as converte solicitadas através do formato TIFF[16] para GIF[17] em tempo real. No entanto, o formato TIFF não é um formato amplamente compreendido pelos navegadores da *Web*, as imagens de página apresentadas ao utilizador são convertidas para o formato GIF uma vez que são duas ou três vezes maiores que a apresentação TIFF (Shaw, Elizabeth J., Blumson, Sarr, Hatcher, Harlan, 1997). Além disso, o MoA oferece aos utilizadores três níveis de resolução de imagem, para acomodar exibições de tamanhos variados e espaçamentos. A pré-computação de todas essas imagens GIF exigiria um maior armazenamento no disco por parte do MoA, foi então que resolveram o problema da seguinte forma:

> We resolve this problem by generating the GIF images only as they are requested. This is facilitated using tif2gif[18], a specialized utility which converts TIFF images to GIF images quickly, but with a limited set of scaling options (Shaw, Elizabeth J., Blumson, Sarr, Hatcher, Harlan, 1997).

O material da coleção MoA é digitalizado a partir do papel original, com os materiais a serem descarregados localmente devido à natureza, por vezes frágil, dos seus itens. Todo o processo de digitalização ficou a cargo da *Northern Micrographics*[19], as imagens foram capturadas em 600dpi no formato de imagem TIFF e compactadas através do Grupo CCITT4[20], utilizado para imagens bitonais (Shaw, Elizabeth J., Blumson, Sarr, Hatcher, Harlan, 1997). O sistema de acesso ao MOA foi desenvolvido de forma progressiva a partir de um sistema de apresentação de imagens de página com informações bibliográficas pesquisáveis para o atual sistema de pesquisa de texto completo.

Esta implementação depende essencialmente de três componentes:

---

[16] TIFF (Tagged Image File Format) - consiste em definir *tags* que descrevem as caraterísticas da imagem; permitem armazenar informações relativas ás dimensões da imagem, ao número de cores e ao tipo de compressão. Disponível em: https://www.techopedia.com/definition/2093/tagged-image-file-format-tiff Acedido em: 9 agosto 2018.

[17] GIF (Graphics Interchange Format) - É um formato de arquivo destinado para ser usado na Web; usa compactação sem perdas e não prejudica a qualidade da imagem, usando cores indexadas. Poderá incluir até 256 cores. Disponível em: https://techterms.com/definition/gif Acedido em: 9 agosto 2018.

[18] Tif2gif – foi desenvolvido na Universidade de Michigan e é usado nas coleções digitais do MOA, escrito por Doug Orr. Disponível em: https://quod.lib.umich.edu/m/moagrp/moa_faq.html Acedido em: 13 agosto 2018.

[19] Northern Micrographics - adapta as novas tecnologias digitais para criar opções mais robustas de preservação, catalogação, recuperação e distribuição da informação. São proprietários dos Softwares ProSeek e PhotoAtlas. Disponível em: http://www.normicro.com/ Acedido em: 9 agosto 2018.

[20] CCIT4 estão relacionados com a transferência de dados. O esquema CCTI4 está destinado para a compressão de bits (ficheiros de texto). Disponível em:https://student.dei.uc.pt/~jsilva/informaticabasica/multimedia/componentes/compressao/ccitt.html Acedido em: 13 agosto 2018.

- Transformação das imagens da página em texto "bruto" e consequentemente a codificação SGML que permite pesquisar o texto;
- Implementação de ferramentas de pesquisa que utilizam o texto codificado para pesquisa e a conversão "just in time" para apresentação;
- Conversão em tempo real das imagens TIFF para o formato GIF para visualização.

Na conversão das imagens da página em texto OCR, este permite que haja pesquisas simples de texto completo dos materiais do MOA. Embora o OCR não seja na sua totalidade perfeito, fornece um acesso significativamente maior do que os bancos de dados bibliográficos simples. De forma a usar o reconhecimento ótico de carateres, na interface de pesquisa, é necessário ter capacidade de reter informações sobre o local da página e a estrutura do documento. Para além disso, o grande número de imagens no projeto exigiu que se automatizasse o processo que poderia vir a ser executado de forma autónoma, apesar dos desafios inerentes a uma coleção com variações significativas no formato, tipo e qualidade de impressão dos materiais originais (Shaw, Elizabeth J., Blumson, Sarr, Hatcher, Harlan, 1997).

O processo automatizado de codificação SGML foi desenvolvido para processar os arquivos de texto não processado para remover carateres *non*-ASCII[21] e limpar o texto. Colocar metadados bibliográficos sobre o documento contido num arquivo previamente preparado através do NMI[22] e inseri-lo num cabeçalho em conformidade com o TEI. Devido às raízes na comunidade de pesquisa humanística, o esquema TEI é impulsionado pelo seu objetivo original de atender às necessidades de pesquisa e, portanto, está comprometido em fornecer o máximo de compreensão, flexibilidade e extensibilidade.

> More specific design goals of the TEI have been that the Guidelines should:
>
> -provide a standard format for data interchange
>
> -provide guidance for encoding of texts in this format
>
> -support the encoding of all kinds of features of all kinds of texts studied by researchers
>
> -be application independent (Burnard, 1995)

[21] ASCII (American Standard Code Information Interchange) é uma codificação de caracteres padrão para a comunicação eletrónica. Estes códigos representam texto em computadores e em equipamentos de telecomunicações. Disponível em: http://zvon.org/other/elisp/Output/SEC525.html Acedido em: 9 agosto 2018

[22] NMI (Non-Maskable Interrupt) consiste na interrupção de softwares e dispositivos de hardware não vitais. Normalmente é usado para verificar se ocorreu algum erro grave e interromper as operações devido a uma falha. Disponível em: https://www.computerhope.com/jargon/n/nmi.htm Consultado em: 9 agosto de 2018.

Relacionar todas as páginas do documento num único arquivo SGML que inclua codificação que marque o conteúdo em divisões maiores no cabeçalho, no corpo e no conteúdo, quebras de página e que retenha referências a imagens sem texto (Shaw, Elizabeth J., Blumson, Sarr, Hatcher, Harlan, 1997).

Os Serviços de Produção de Bibliotecas Digitais da Universidade de Michigan através da *Humanities Text Initiative*[23] (HTI), têm uma considerável experiência com textos codificados em SGML e o mecanismo de pesquisa exibe dinamicamente informações na *web*. O texto codificado em SGML fornece informações estruturadas para pesquisas em campo que podem exibir e reter informações sobre o contexto texto (Shaw, Elizabeth J., Blumson, Sarr, Hatcher, Harlan, 1997). Permite identificar informações bibliográficas sobre o documento, o número de "acessos" em páginas individuais e informações que utilizam o sistema de apresentação de imagens.

Nas funções de pesquisa e navegação é utilizado um *script* CGI[24], que estende os modelos desenvolvidos a partir do HTI e gere informações do formulário, resolve a pesquisa no idioma pesquisado através do mecanismo de *Open Text*.

> Using two modules (developed at the University of Michigan) that manage the interaction and its results, the search is handed off to the search engine to search the indexed SGML. Results are passed back to the CGI script. The CGI script filters and displays the resulting data. The first results screen provides a list of documents matching the search query. If the search has been a full text search it also displays the number of hits in that document text. (Shaw, Elizabeth J., Blumson, Sarr, Hatcher, Harlan, 1997).

Quanto à exibição de página, a interface está dividida em dois quadros, o superior que contém a imagem da página, enquanto o quadro inferior contém botões de navegação.

> To minimize interaction between the viewer and the search engine, page navigation uses the page numbering conventions. The page images are stored in files with names of the form XXXXYYYY.tif, where XXXX is the ordinal number of the page in the sequence of bound pages (as shown in parenthesss in the example) and YYYY is the printed page number that appears on the page (outside the parentheses). The

[23] HTI (Humanities Text Initiative) é uma organização abrangente para a criação, entrega e manutenção de textos eletrónicos, bem como mecanismo para promover as capacidades da comunidade de uma biblioteca online. Disponível em: http://www.hti.umich.edu/ Acedido em: 13 agosto 2018.

[24] CGI (Common Gateway Interface) define um padrão em que as informações podem ser passadas para e do navegador e servidor. É uma forma mais simplificada de colocar um programa online, para qual os utilizadores podem enviar dados. Disponível em: https://users.cs.cf.ac.uk/Dave.Marshall/PERL/node187.html Acedido em: 13 agosto de 2018.

> first digit of YYYY is replaced by an "r" if the page number is a Roman numeral text. (Shaw, Elizabeth J., Blumson, Sarr, Hatcher, Harlan, 1997).

Usando isto, a página anterior ou a próxima, podem ser identificadas subtraindo ou adicionando, uma da parte “XXXX” do nome do arquivo de página atual e localizando o arquivo correspondente. Existe outro método de navegação, o menu *go to* que permite ao utilizador pular diretamente para as páginas de que tem interesse, como a página de título ou o sumário. Uma vez que estas páginas podem ser identificadas, ou até mesmo ser verificada a sua existência, apenas por referência aos dados SGML. O menu *view as*, disponibiliza alternativas para o tamanho/resolução da imagem, permitindo que o utilizador escolha o tamanho de exibição ideal com base na resolução da sua tela e noutros aspetos de visualização texto (Shaw, Elizabeth J., Blumson, Sarr, Hatcher, Harlan, 1997).

CAP 3 PROPOSTA DE MODELO PARA A CONSTRUÇÃO DE UMA HEMEROTECA DIGITAL

Neste capítulo irão estar representados os requisitos que assistem numa orientação para criação de projetos de Hemerotecas Digitais em bibliotecas públicas ou outro tipo de instituição que pretender fazer uso dele.

No intuito de servir de guia, os requisitos funcionais que são apresentados no Quadro 1, foram recolhidos com base nos Projetos suprarreferidos, isto é, dos itens e das referências que sustentaram cada um deles.

Embora os projetos estudados estejam direcionados para bibliotecas específicas ou instituições, esta tabela/modelo foi criada com o objetivo de servir as bibliotecas públicas. É sabido que estas instituições são muitas vezes dependentes de fundos municipais para criarem um projeto e são também as mais importantes para a construção de conhecimento para a sua comunidade, daí estabelecerem-se pontes entre importantes projetos internacionais e criar este projeto. Muitas bibliotecas públicas em Portugal possuem um acervo riquíssimo de jornais e essa é uma das justificações para a importância de o disponibilizar para toda a comunidade incluindo investigadores de todo o mundo. Para além de enaltecer a cultura local e/ou da instituição, a digitalização é um dos principais meios de conservação e preservação dos jornais.

É importante ter em conta, antes de avançar com os pressupostos, que a escolha dos requisitos preza pela importância em termos técnicos e económicos: quanto mais dispendiosa é uma tarefa, menos "obrigatoriedade" tem em ser utilizada.

A Digitalização consiste no processo pelo qual um determinado dado (imagem, som, texto) é convertido para o formato de dígito binário para ser processado por um computador (Martins, 2008).

> A digitalização está direcionada para a preservação, acesso e difusão de um acervo, contudo, o produto da conversão não será igual ao original e não substituiu de forma alguma, a preservação do original (Giordano, 2016).

Existem alguns agentes capazes de deteriorar as coleções, como é o caso da sua composição química, a humidade, a exposição à luz, o manuseamento e os insetos. Assim, a digitalização é uma forma de preservar o documento original. Outro dos objetivos da digitalização é tornar o acervo disponível para todos os utilizadores, sem que seja necessária a deslocação física à instituição. É importante ter em conta a tecnologia a ser usada no processo de digitalização de documentos, os *softwares* e os *hardwares* (neste caso os scanners) têm um papel fundamental no desenrolar da tarefa: a qualidade da imagem, o formato em que é guardada (JPG, PNG ou GIF), o tratamento da imagem e a salvaguarda em bases de dados.

O Processamento é o requisito que suporta a entrada em ação das funcionalidades dos *softwares* OCR e OLR (escolhidos pela instituição segundo critérios económicos, por exemplo) e à transformação da imagem digital em metadados. A utilização de metadados na organização eletrónica de recursos, vem ao encontro da necessidade crescente de descobrir e disponibilizar informações na internet (Alves e Souza, 2009). A catalogação e descrição dos documentos é feita a partir da análise do documento e inserida na base de dados de forma a ser facilmente encontrada e recuperada. Como vamos observar mais à frente, o Processamento é o trâmite que se segue à Digitalização e o que antecede os propósitos da Publicação, estando responsável pelas análises ótica de carateres e de *layout* e à sua intervenção de correção (se for caso disso). O último ponto identificado corresponde ao requisito da Publicação. Este é o último uma vez que lhe corresponde a tarefa de transformar os metadados em números binários para a sua aplicação em ambiente *web*. É neste campo que são decididos os resultados que se irão obter a partir das pesquisas dos utilizadores, se simples ou mais elaborados. É neste espaço fundamental que se constrói todo o ambiente web para que o utilizador tenha acesso aos conteúdos de forma interativa, simples e satisfatória. Opções como a ferramenta do zoom, nas imagens, é um ponto a ter em consideração como vamos poder discutir mais abaixo.

3.1 Objetivos

3.1.1 Objetivo geral

O modelo que a seguir se apresenta foi construído com base nos projetos descritos anteriormente e correspondem a um guia para a construção de uma Hemeroteca Digital. Os elementos foram categorizados sob três pontos de vista: os obrigatórios, os recomendados e os facultativos. Estes três blocos correspondem a outras três grandes etapas para a construção de uma Hemeroteca Digital: a Digitalização, o Processamento e a Publicação.

3.1.2 Objetivo específico

Num primeiro plano, é necessário ter em conta aquilo que a instituição pretende fazer. Ao digitalizar-se e disponibilizar em sítio na Web estamos a estabelecer a consulta gratuita de um acervo que integra uma instituição e com isso o fortalecimento da própria identidade perante a comunidade ou perante o espetador global; ao navegar-se pelas páginas dos jornais de forma interativa, temos como consequência positiva a promoção da cultura local e/ou institucional.

3.2 Especificações

O modelo apresentado foi criado para corresponder a tarefas que serão realizadas após a instituição decidir quais os documentos a digitalizar, se tem recursos disponíveis para o fazer e qual o tempo de execução das tarefas (esse estará assegurando segundo o número de colaboradores que a instituição detenha e o tamanho do acervo dos jornais, de modo a distribuir as funções.).

Os pressupostos que aqui se reúnem surgiram após a análise de três distintos projetos, o NPND, a Europeana e o *Making of America*. O conjunto de dados apresentado pelos documentos partilhados pelos vários projetos, são recursos valiosos para iniciativas de digitalização. A sua usabilidade é ampla e dispõem de um gama representativa de conteúdos, bem como a verdade detalhada que sublinha a sua singularidade.

Através do projeto da Europeana, foi possível desenvolver e partilhar as melhores práticas para a digitalização do conteúdo dos jornais. O projeto concentra-se em (Reilly, 2013):

- Uso de métodos de refinamento para OLR, OCR E NER;
- Reconhecimento de classe para aprimorar funcionalidades de pesquisa e apresentação para os utilizadores da Europeana;
- Avaliação de qualidade para tecnologias de refinamento automático;
- Transformação de metadados locais para o modelo de dados Europeana (EDM);
- Padronização dos metadados.

O site da Europeana permite que os utilizadores façam o *download* de conjuntos de uma ou mais imagens de documentos originais de alta qualidade (cores até dois níveis) e os arquivos de realidade associados, selecionando uma quantidade de documentos de uma lista de navegação ou de resultados de pesquisa. Os itens selecionados são agrupados num arquivo ZIP em tempo real (Antonacopoulos, Bridson, Papadopoulos, & Pletschacher, 2009).

O projeto MoA levou a que, antes de começar qualquer projeto de digitalização se considerassem algumas questões (Hagedorn, Kate, 2016):

1. Dimensão da coleção

O importante é ter uma ideia, senão saber ao certo quantos itens na realidade temos na nossa coleção. Quanto mais precisa for a contagem, mais acertado será o nosso planeamento. A

equipa que guiará o projeto pode começar com uma estimativa aproximada e refinar a partir daí, contudo, será sempre necessária uma contagem precisa aquando da digitalização.

2. Tipologia de documentos

As coleções podem ter vários formatos – texto, mapas, fotografias, correspondência, áudio, vídeo, etc. O mais importante é ser o mais específico possível.

3. Disponibilidade dos documentos noutras instituições

Se na nossa coleção existirem documentos que estejam já disponíveis noutras instituições em formato digital, é possível diminuir os custos de digitalização da coleção. Investigar se outras instituições digitalizaram os mesmos documentos que temos na nossa coleção é fundamental. Assim, saberemos se devemos gastar mais tempo e recursos na digitalização.

4. Caraterísticas da coleção: títulos vivos ou mortos

A coleção física é definida em tamanho ou o novo material será adicionado ao longo do tempo? Se a coleção estiver a crescer, é importante dar indicações de com que frequência o material é adicionado e quando é que o novo material irá ser digitalizado e adicionado à coleção digital ou não.

5. Duplicados

É importante verificar se realmente existem documentos duplicados. Ao se verificar se existem documentos duplicados estaremos a poupar tempo e recursos, a não ser que exista uma necessidade absoluta de apresentar o mesmo conteúdo mais que uma vez. A recomendação é de que se deve percorrer a coleção e observar onde existem duplicados. Se um dos duplicados for melhor que o outro, há que o escolher por esse mesmo fator. Os projetos de digitalização podem, de igual modo, ser um bom momento para rever o que está na coleção e eliminar material que não é necessário.

6. Verificação de itens já digitalizados

Se algum dos documentos já estiver digitalizado ou esteja em formato digital, é necessário rever os ficheiros. Precisamos de ter a certeza de que eles não se voltem a digitalizar. Nos formatos digitalizados, os itens digitais podem ser proprietários ou de acesso aberto, ou se a conversão para outro formato digital é necessária.

7. Processo de digitalização: na instituição ou entregue a uma empresa

O envio de materiais para um fornecedor geralmente resulta numa resposta mais rápida, mas alguns documentos podem não estar em estado de sair da instituição. Nesse caso, é necessário saber quais os mais debilitados. O microfilme é um exemplo de material que poderá ser apenas manuseado por especialistas.

8. Esclarecer o objetivo da criação de uma Hemeroteca Digital de Jornais

É de considerar qual a forma em que os utilizadores poderão usar a coleção de jornais digitais. O que fará com que se sintam atraídos para aceder à Hemeroteca? É importante tecer considerações sobre quais as opções aplicáveis ao manuseio desta ferramenta. O que lhes irá ser apresentado e de que forma estará feito é fulcral para o funcionamento da hemeroteca.

9. Identificar as funções consideradas relevantes para os utilizadores:

- Pesquisar texto;

- Navegar como um álbum de fotos online;

- Fazer download do conteúdo ou do próprio documento;

- Estender o uso para além da intenção original;

- Fornecer maior acessibilidade a diferentes grupos de utilizadores.

De igual modo se deverá questionar qual o valor que realmente a coleção a digitalizar oferecerá. Se realmente queremos construir uma hemeroteca digital de jornais, é necessário faze ruma revisão dos materiais que mais se adequam à missão do projeto, para que possamos oferecer acesso o acesso à informação e à sua disseminação. Desses valores, estão incluídos:

- Valor administrativo;

- Valor como um artefacto;

- Valor a nível educacional;

- Valor como evidência;

- Valor monetário substituído pelo conteúdo digital.

Quando falamos em opções de financiamento para a digitalização, a maioria das instituições não financia a real fase da digitalização do projeto, no entanto, todos os aspetos do projeto

podem ser apoiados financeiramente ou ter custos compensados por financiamento externo. Todo esse processo inclui:

- Preparação dos materiais originais para digitalização;

- Desenvolvimento de recursos, conforme necessário, na interface para a coleção;

- Conservação dos objetos físicos antes ou depois da digitalização;

- Manutenção a longo prazo dos arquivos digitalizados, uma vez que os servidores têm custo de manutenção ao longo dos anos.

Se os fundos da instituição não estiverem disponíveis ou não forem suficientes, a próxima melhor opção é através de subsídios. A maioria das doações de legalização proporcionará dinheiro para suportar certos custos e exigirá que a instituição suporte o restante. Haverá alguns subsídios que cobrirão apenas aspetos específicos, como a transformação em metadados e a catalogação. Independentemente disso, o financiamento de qualquer aspeto é melhor do que não ter nenhum (Hagedorn, Kate, 2016).

Ao digitalizar-se, procede-se à transferência do conteúdo intelectual dos documentos para o suporte digital. Segue-se a preocupação com o tratamento da imagem e do texto e, como tal, é importante definir os destinos e, em função destes, a qualidade da resolução da reprodução digital. Como é de se esperar, a qualidade está depende essencialmente dos programas de tratamento de imagem utilizados, dos equipamentos de captura de imagens (*scanner*) e também a sensibilidade de quem procede com a atividade.

Depois do tratamento do texto e da imagem, segue-se a transferência desta informação, ou seja, procede-se ao "entrelaçamento" do suporte digital com os respetivos registos bibliográficos (Matos, 2000).

3.3 Análise dos requisitos funcionais

De caracter Obrigatório na Digitalização

O primeiro ponto obrigatório para proceder à digitalização é garantir a existência de recursos humanos suficientes para distribuir as tarefas tais como: analisar o acervo, distribuir competências, priorizar os documentos a digitalizar e verificar o armazenamento disponível. Ao digitalizar-se, os ficheiros têm de obrigatoriamente ficar armazenados na memória dos computadores da instituição. Verificar se existe espaço de memória suficiente vai impedir que, enquanto se digitaliza, o trabalho esteja a ser feito em vão.

Ter equipamento para trabalhar é de igual modo um ponto obrigatório, visto que sem ele, não é possível proceder a nenhuma tarefa. *Scanner* ou *scanner* planetário serão alguns dos equipamentos fulcrais neste processo. Existem centenas de marcas e preços, a instituição deverá ter em conta a qualidade com que quer que seja feita a digitalização, o tamanho da resolução após a digitalização (varia de 300 dpi's a 600), qual o formato do documento após a digitalização (PDF, JPEG, TIF, etc.), verificar se o tamanho da base é suficiente para digitalizar toda a área do jornal e equilibrar estas variantes consoante o orçamento disponível para a compra. Algumas marcas de scanner disponíveis no mercado:

- Epson;
- HP;
- Xerox;
- Canon.

Após a aquisição de todos os materiais necessários para proceder à digitalização e análise geral da coleção vem a fase da seleção. Procede-se à seleção de prioridades para digitalizar como afirma (Giordano, 2016):

> (…) orienta-se pelo estado de conservação de determinado item, pelo tipo de material disponível, pela procura e uso do item no acervo e também por influência de patrocinadores dispostos a contribuir para a digitalização de uma coleção específica.

É tido como obrigatório que, a digitalização seja frente e verso das páginas de jornais e que sejam digitalizadas todas as páginas, de forma a evitar lacunas de leitura e compreensão.

De caracter Recomendado na Digitalização

O técnico responsável pela digitalização deve tomar medidas de precaução aquando do manuseamento dos jornais, especialmente aqueles que apresentam um estado mais debilitado. Com as luvas, máscara e bata, o técnico procede à preparação dos documentos: observa as folhas, desvincado os cantos, retirando agrafos ou clipes que estejam anexados e faz uma limpeza.

De carater Facultativo na Digitalização

Como a aquisição dos equipamentos, como os scanners dependem essencialmente dos recursos monetários da instituição, é facultativo que as câmaras que captam as imagens, sejam de alta precisão.

O laboratório onde seja feita o manuseamento e a digitalização dos jornais terá de reunir condições a nível de humidade relativa, temperatura e luz. A disposição do equipamento é também algo facultativo no sentido em que, depende dos fatores referidos anteriormente.
Após a digitalização dos documentos que a equipa selecionou, estes são transportados para o computador e não existe recomendação para o formato da imagem. Poderá ser em formato PDF, TIFF, GIF, PNG ou JPEG (ou JPG). No entanto, as opções de extensão PDF e TIFF são mais sofisticadas do que o JPEG, uma vez que permitem informações a nível de direitos de autor e indicam as dimensões do ficheiro. O JPEG é de fácil compressão, mas isso não significa qualidade quando a imagem é transportada para as páginas da internet. O formato GIF é uma opção melhor que o JPG, quando, as imagens que vamos inserir na internet não necessitem de muita qualidade. O nível de compressão é maior e oferece cores mais vivas.

De carater Obrigatório no Processamento

Os resultados de texto que completam o processamento não são diretamente entregues aos utilizadores, uma vez que o principal objetivo, por detrás da conversão de texto é, fornecer aos utilizadores a opção de pesquisa do texto completo. Portanto, a conversão de texto não visa obter alta precisão de OCR, mas sim, a criação de um arquivo que corresponda às expectativas para a pesquisa do texto completo por parte dos utilizadores.
Quando associado a bases de dados bibliográficos ou sujeito à intervenção do OCR oferece-nos extraordinárias capacidades na difusão e no acesso à informação, facilitando e alargando esse acesso de uma forma que podemos considerar revolucionária, tomando a digitalização numa forma privilegiada para a comunicação e para a recuperação da informação. Sobre as vantagens da utilização do OCR, esta ferramenta abre-nos o acesso a um outro caminho na recuperação da informação, isto é, permite-nos, com o auxílio do *thesaurus* que é construído, uma pesquisa integral ou difusa, que oferece uma grande flexibilidade, facultando-nos múltiplas aproximações à palavra que está a ser utilizada na pesquisa e, simultaneamente, permitindo-nos avaliar o comportamento do motor de pesquisa que lhe está associado.
Após a digitalização de uma página de jornal, poderá ser efetuada uma análise de *layout* para reconhecer os elementos padrão de um artigo com títulos, legendas, colunas, imagens e nomes de autores. Para além disto, o software também analisa a ordem desses elementos padrão. Como verificamos com o Projeto da Europeana, é necessário analisar a imagem digitalizada e saber o que temos presente, fazendo com que seja mais fácil identificar texto para o *software* OCR poder entrar em ação.

O processamento de reconhecimento ótico de carateres entra em ação nas partes bibliográficas do artigo digitalizado. O resultado do processamento é, geralmente, satisfatório, pelo menos no que respeita aos títulos e legendas. Se for detetado algum erro, será apresentada uma ferramenta que permite corrigir o texto de forma imediata. Para poupar algum tempo, a tarefa deverá corresponder ao fluxo de trabalho orientado, isto vai fazer com que o processamento OCR e a exportação de dados final fiquem completamente automatizados.

> OCR software creates machine-readable text from scanned page images and permits full- text searching of the contents of newspaper pages. Bounding-box data relates words to their position on the image (Digital & Program, 2009).

O método para o reconhecimento ótico de carateres geralmente precisa de ser aperfeiçoado para reconhecer novas fontes e símbolos:

> Training data can be either synthesised or extracted from document images (ground truth). Especially in the case of historical documents a synthesis is usually not feasible because font descriptions are not available. (Clausner, Pletschacher, & Antonacopoulos, 2014)

A instalação de um software OCR é fundamental uma vez que vai transformar imagens em arquivos de texto editáveis. O OCR serve essencialmente para editar erros que existam ou aprimorar informações, tal como os títulos dos jornais, datas, localizações e nomes. Para software foi escolhido o METS/ALTO que armazena informações de layout e o OCR reconhece o texto de páginas de qualquer formato, os jornais estão obviamente incluídos.

> ALTO is a standardized XML format to store layout and content information. It is designed to be used as an extension schema for use with the Library of Congress: Metadata Encoding and Transmission Schema (METS) XML Schema, where METS provides metadata and structural information while ALTO contains content and physical information.

Após a digitalização é indispensável nomear os ficheiros, de forma a não se perder e a serem facilmente identificados. É certo que esta prática varie de instituição para instituição, elegendo assim o que lhes for mais conveniente: a sigla da instituição, numeração árabe, numeração romana e até a mistura dos dois. A criação de um banco de dados surge com a classificação, indexação e catalogação dos documentos digitalizados, criando assim metadados que ficam à disposição nos servidores existentes (Keffer,2012). O banco de dados contém metadados relacionados com as condições encontradas em documentos e necessidades de manutenção que podem ser de interesse para os utilizadores. As informações que devem ficar armazenadas no banco de dados são:

1. Informação bibliográfica: título, autor, data, publicação, etc.;
2. Informações de imagem: scanner utilizado, resolução e escala;

3. Recursos de layout: presença de imagens ou gráficos, variedades de tamanho da fonte e número de colunas;
4. Informação administrativa: detetor de direitos de autor.

> Os metadados possibilitaram registos das principais informações sobre cada exemplar: título; autorias (editores, redatores, gerentes); idioma; país e cidade da publicação; entidade responsável pela publicação; data (dia, mês e ano); volume e número da edição; número de páginas do exemplar; assuntos; e notas (para destaque de notícias de grande interesse) (Renato & Galvão, 2013).

A catalogação é uma parte essencial para a criação de conteúdo online, sobre o protocolo e o nível de detalhe da descrição e os idiomas presentes nos documentos. A profundidade da descrição deve ser equilibrada, deve ter-se em conta o utilizador e o potencial utilizador.

De carater Recomendado no Processamento

Continuando a falar da relevância do reconhecimento ótico de carateres, não é considerado obrigatório que se faça a sua correção. É sabido que após a digitalização de uma página de jornal e após a intervenção do OCR, este não atue na perfeição. Deste modo, é recomendado que se faça uma correção no sentido de aperfeiçoar o texto que ficará disponível para o utilizador.

Num nível futuro, podemos imaginar um portal de jornais que, para além de nos apresentar jornais digitalizados, também forneceria o serviço chamado de *magic glasses*. Sem estes óculos mágicos, os utilizadores verão apenas o texto do OCR com todas as suas falhas, erros, palavras irreconhecíveis e lugares não identificados.

Com os *magic glasses*, podem ser feitas as seguintes alterações em texto OCR (Zogla, 2014) :

- Correção de erros;
- Correção de linguagem obsoleta (modernização);
- Explicação de entidades nomeadas. Estas explicações podem ser recuperadas de fontes externas, como por exemplo, da Wikipédia;
- Explicação de expressões idiomáticas;
- Substituição de endereços antigos por endereços equivalentes e modernos;
- Tradução do texto para a língua original do utilizador.

Existem três cenários para implementar o serviço dos *magic glasses* (Zogla, 2014):

> 1. Pre-process all text with the magic glasses service once and later, when activated, it would simply retrieve the enhanced version of particular piece of text;

2. Crowd-source adding all enhancements, but this even in the best-case scenario, would provide improvements only to some parts of text archive, while other parts would remain unchanged;

3. Run the magic glasses service on-the-fly.

É recomendado que os diversos documentos digitalizados sejam armazenados em servidores temporários, de forma a garantir a segurança do trabalho efetuado. Se por algum inconveniente os jornais digitalizados deixarem de estar disponíveis, uma fonte segura de os recuperar será certamente no armazenamento, sendo a c*loud* o mais indicado.

Caso as imagens dos jornais sofram edição de forma a melhorar a qualidade da imagem como, a cor, a nitidez, o contraste ou o brilho é sugerido que esta informação seja indicada nos campos adequados. Tal como uma escala. Isto é, as escalas servem como fonte de guia para perceber qual o tamanho real do jornal, é recomendado que esteja uma escala projetada na imagem para que o utilizador fique com a informação real do documento.

De carater Facultativo no Processamento

É facultativo o uso de marcas d'água com a sigla da instituição (ou outra qualquer) após a digitalização das páginas dos jornais. As marcas d'água têm como objetivo identificar o proprietário intelectual do documento, de forma a combater o uso ilícito das imagens digitais. É importante perceber que a marca d'água não deverá interferir com a leitura do documento disponibilizado. Contudo, estes pressupostos são debatidos perante os interesses da instituição que estiver a digitalizar.

O *software* NER é outro componente que apresenta um carater opcional, com o objetivo de recolher informações acerca de entidades nomeadas como, nomes de pessoas, nomes de organizações, valores monetários, expressões ou localizações. Foi colocado neste campo uma vez que não é elementar a recolha deste género de informações.

De carater Obrigatório na Publicação

Quando se decide publicar o trabalho feito até aqui, é importante que reflitam sobre a importância do reconhecimento ótico de carateres e que, divulgá-lo ao utilizador poderá ser útil. A questão de disponibilizar os resultados fará sentido apenas se este for corrigido, de forma a contribuir para o enriquecimento de quem está a consultar.

Os formatos de navegação devem ter um carater simples, claro, eficiente. O que o utilizador mais preza é uma interface de pesquisa simples, chamativa e interativa. O objetivo é permitir

que o utilizador tenha acesso a todas as ferramentas e saiba para que elas realmente funcionam. Manter um perfil conciso vai fazer com que as pesquisas se tornem mais eficazes e desperta o interesse de quem estiver a consultar. Nisto inclui a forma ágil com que as ferramentas desempenham o seu trabalho, as cores que se utilizam. É importante manter um perfil adaptado aos utilizadores e que os chame a atenção.

O conteúdo criado após a digitalização, o tratamento das imagens e a transformação em metadados, pode ser acedido em quatro níveis:

- Listas de miniaturas (navegação através dos resultados de pesquisa);
- Detalhes do documento (imagem de visualização, metadados e *links* que sejam úteis);
- Pré-visualização das imagens;

Quando o utilizador, após obter os resultados de pesquisa ou mesmo durante a navegação faz um clique numa miniatura, uma página de informações intermediárias tem de ser exibida contendo uma miniatura maior, para uma visualização mais próxima, bem como metadados a nível do documento. Será obtida uma visualização de resolução completa, após o utilizador requerer uma visão mais detalhada do documento que pesquisou.

De carater Recomendado na Publicação

É indicado que os resultados sejam exibidos a partir de pesquisas simples. Este género de pesquisa é muito intuitivo e fácil de usar, basta apenas inserir as palavras-chave daquilo que pretende recolher. Palavras simples como o título do jornal, datas ou nomes devem poder ser utilizadas.

Como já foi referido, é importante escolher o formato da imagem após a digitalização, no entanto existe a recomendação de que a instituição as deverá disponibilizar com boa resolução. Tem que se ter em conta o formato da imagem, uma vez que algumas delas têm a capacidade de compressão sem perder a qualidade, enquanto outras não conseguem apresentar as mesmas caraterísticas. Se os jornais que digitalizamos têm muitas fotografias ou imagens, este é um procedimento que se deve ter em conta. A ferramenta do zoom, por exemplo, vai ter mais funcionalidade quando as imagens oferecerem melhor qualidade. Isto vai dar ao utilizador maior experiência na exploração dos recursos que encontrou.

Não existe qualquer preceito em ter de se instalar algum tipo de *software* para os utilizadores conseguirem aceder aos jornais digitalizados. As imagens são disponibilizadas a partir do formato escolhido.

De carater Facultativo na Publicação

Continuando a falar sobre o género de pesquisa, optamos que o formato de pesquisa booleana não seja vinculativo. A pesquisa booleana está representada através de operadores que exercem algum tipo de relação lógica entre os termos a pesquisar: AND, NOT e OR.

Caso o formato escolhido seja o PDF, por exemplo, há que se considerar a hipótese de a plataforma, ter capacidade de *download* das imagens originais. Todos os procedimentos implícitos, estão sob o cumprimento dos direitos de autor.

Este conjunto de princípios constitui a base da proposta de modelo sumariada no quadro seguinte.

QUADRO 1. PROPOSTA DE MODELO PARA A CONSTRUÇÃO DE UMA HEMEROTECA DIGITAL

	Digitalização	**Processamento**	**Publicação**
Obrigatório	Recursos humanos	*Software* OCR	Publicar o que foi extraído do OCR
	Verificar armazenamento disponível	*Software* OLR	Navegação simples
	Equipamento para digitalizar: *scanner*, por ex.	Metadados bibliográficos	Listas de miniaturas
	Digitalizar frente e verso		Pré-visualização das imagens
	Fazer *scanner* de todas as páginas dos jornais		
Recomendado	Calcular o tempo da tarefa a ser executada	Correção manual OCR	Campo de pesquisa simples
	O equipamento deve ter o tamanho adequado aos docs.	Armazenamento em servidores temporários	Ferramenta de zoom
	O equipamento deve ter base própria	Editar imagens	Não instalação de *software*
	Atender ao manuseio de jornais antigos	Associar escala	
	Preparação dos documentos		
	Usar equipamento necessário para manuseamento		
Facultativo	Câmaras do equipamento de alta precisão	Uso marcas d'água	Pesquisa booleana
	Disposição do equipamento	*Software* NER	Ter capacidade de *download*
	Condições do laboratório		
	Formato das imagens após digitalização		

CONCLUSÃO

O jornal desempenha um papel fundamental para o conhecimento e para a história, através das suas colunas informativas, das imagens e das fotografias. Receber, organizar, identificar e guardar a memória do país e /ou da cidade onde está inserida, é o papel desempenhado por uma hemeroteca.

Em virtude dos aspetos mencionados, é de extrema importância ter em conta a digitalização de documentos, como os jornais, para a sua conservação e preservação. Uma vez que os suportes de informação se diversificam, a disseminação da informação torna-se complexa no sentido de utilização de ferramentas que proporcionem o acesso à informação em ambiente digitais. As bibliotecas públicas, hoje em dia, estão cada vez mais concentradas em resolver questões do âmbito das práticas para a conservação, preservação e disseminação das suas coleções. É importante existirem medidas de preservação do acervo, a forma como é manuseado, o local onde está guardado (combinar temperatura e humidade relativa) e inventariar cada um deles de modo a não haver uma perda maior. A preservação digital é reconhecida pela sua crescente importância e técnicas digitais de armazenamento e transporte da informação, como é o exemplo da digitalização. A proposta de digitalização de jornais é uma das grandes medidas que poderá ajudar a conservar a memória e a informação. De caráter moroso e dispendioso, a digitalização é um ato consciente de preservação. Este ato, é pensado a partir da avaliação para o estabelecimento de critérios de análise que permitem adaptar as funcionalidades da digitalização à preservação e das necessidades dos utilizadores e de futuros utilizadores. A responsabilidade desta tomada de decisão é essencialmente dos bibliotecários ou responsáveis de cada instituição, pelo que, quando se intervém de alguma forma, seja a digitalização, a catalogação ou a publicação através da web, são obrigados a garantir o acesso a essas coleções. Conseguimos perceber que, pelo uso do sistema *handle*, vai facilitar que o documento seja encontrado mais facilmente através do seu endereço na *cloud*. Para além de todos os aspetos positivos já mencionados acerca da digitalização de jornais, existem outros tais como, a facilidade de aceder a um jornal a partir do computador de casa e navegar por entre as suas páginas. A digitalização é igualmente uma forma de "manusear" virtualmente coleções antigas, sem que haja danificação do material. Existem de igual modo instituições que, embora não tenham digitalizado o seu acervo, desenvolveram bases de dados para que os interessados possam

consultar o catálogo antes de se dirigirem à instituição. Conclui-se então que terá de existir uma sensibilidade formal na abordagem da digitalização, com o objetivo de garantir o acesso à informação e à salvaguarda dos documentos físicos.

Os três pontos fulcrais (digitalização, processamento e publicação), representados através do modelo para o conjunto de orientações para a criação de uma hemeroteca digital foram essenciais para se estabelecerem critérios de aplicabilidade (obrigatório, recomendado e facultativo). Foi a partir do estudo dos projetos de digitalização do NDNP que se estabeleceu a importância da consistência técnica para reproduzir melhores práticas para ir de encontro às necessidades de preservação. A utilização de mapas para apontar a origem geográfica dos jornais, a colaboração de vários utilizadores e o uso do microfilme foram aspetos relevantes para o NDNP. Com parceria com a Biblioteca do Congresso, este projeto tem uma lista de prémios que financiam outros projetos de digitalização em toda a área dos EUA. O projeto da Europeana é aquele que mais conteúdo bibliográfico reproduz. Conta com o apoio de diversas bibliotecas nacionais europeias e foi através dele que percebemos que é importante considerar a realidade de criação de uma identidade a partir de diversas coleções, seja em contexto histórico, cultural e social. Já o último projeto estudado, MoA, mostrou-nos que usam resultados não processados, nem corrigidos após o reconhecimento ótico de carateres e que a digitalização é feita a partir do papel original dos documentos.

Após o estudo intensivo dos projetos acima identificados, foram recolhidos os elementos considerados essenciais para a construção do modelo. Assim, entendemos que o importante é que as instituições transponham barreiras técnicas e que se adaptem à manipulação de programas computacionais. É necessário reconhecer as interconexões entre documentos e estabelecer roteiros de investigação, reconhecer falhas e convertê-las no produto final: que os utilizadores tenham acesso aos documentos digitalizados em ambiente digital. Todo o processamento centrado nesta proposta de modelo para a construção de uma hemeroteca digital reuniu aspetos que fossem considerados adaptáveis, no entanto, apresentam algumas fragilidades no que concerne a custos monetários: custos de mão-de-obra, custo de *hardwares* e *softwares* e custos de manutenção e produção de metadados. No entanto, é possível combater certas tenacidades através de apoios de fundos comunitários e institucionais. É importante consciencializarmo-nos da importância deste género de projetos e aplicá-los. Este guia de boas práticas poderá considerar-se como uma base de lançamento para futuros projetos

que podem ser adaptados aos objetivos e às circunstâncias de cada instituição.

REFERÊNCIAS BIBLIOGRÁFICAS

ABBYY Technology Portal. (2017). *Object Model FineReader Engine 11*. [Em linha] [Consult. 6 agosto 2018]. Disponível em: https://abbyy.technology/en:products:fre:win:v11:object-model Acedido em: 6 agosto 2018.

Adcock, E. P. (1998). IFLA - *Principles for the Care and Handling of Library Material*. *Preservation*, *1*(one), 1–72. [Em linha] [Consult. 16 junho 2018]. Disponível em: http://archive.ifla.org/VI/4/news/pchlm.pdfhttp://archive.ifla.org/VI/4/news/pchlm.pdf

ALTO. (2016). Technical Metadata for Layout and Text Objects. [Em linha] [Consult. 21 agosto 2018]. Disponível em: https://www.loc.gov/standards/alto/description.html

Alves, M. D. D. R., & Souza, M. I. F. (2007). *Estudo de correspondência de elementos metadados: DUBLIN CORE e MARC 21*. *Revista Digital de Biblioteconomia e Ciência da Informação*, *4*(2), 20-38. [Em linha] [Consult. 22 junho 2018]. Disponível em: http://www.brapci.inf.br/index.php/article/view/0000007463/bd666743664eed387696a4ac45d0310e/

ANACOM. (2018). Portugal no Conselho da UIT. [Em linha] [Consult. 20 junho 2018] Disponível em: https://www.anacom.pt/render.jsp?contentId=65196

Arellano, M. (2004). *Preservação de documentos digitais. 15–27*. [Em linha] [Consult. 22 junho 2018] Disponível em: https://www.doi.org/10.1590/S0100-19652004000200002

Assis Boeres de, S. A., & Faria, A. C. C. (2012). *A preservação digital na biblioteca central da Universidade de Brasília*. [Em linha] [Consult. 22 junho 2018] Disponível em: http://www.revista.ibict.br/ciinf/article/view/1363

Assunção, R. V. D. (2011). *Biblioteca digital: uma abordagem conceitual*. [Em linha] [Consult. 22 junho 2018] Disponível em: http://www.rabci.org/rabci/sites/default/files/BIBLIOTECA%20DIGITAL%20uma%20abordagem%20conceitual.pdf

Boddie, Stefan. (2014). *What is METS/ALTO?* [Em linha] [Consult. 24 julho 2018]. Disponível em: https://www.veridiansoftware.com/knowledge-base/metsalto/

Breuel, T. (2007). *The hOCR microformat for OCR workflow and results. In icdar* (pp. 1063-1067). [Em linha] [Consult. 24 agosto 2018] Disponível em:https://www.computer.org/csdl/proceedings/icdar/2007/2822/02/28221063- abs.html

Burnard, L., & Sperberg-McQueen, C. M. (1995). *TEI lite: An introduction to text encoding for interchange*. [Em linha] [Consult. 22 agosto 2018] Disponível em: http://www.tei-c.org/Vault/P4/Lite/teiu5_en.pdf

Calvanese, D. (2000). *Building a digital library of newspaper clippings: The LAURIN Project*. [Em linha] [Consult. 24 junho 2018] Disponível em: https://www.researchgate.net/publication/220716010_Building_a_Digital_Library_of_Newspaper_Clippings_The_Laurin_Project

Carvalho, B. L. P. de (2016). *Digitalização de jornais: uma reflexão sobre desafios e melhores práticas. Acervo, 29* (2 jul-dez), 89-102. [Em linha] [Consult. 20 junho 2018] Disponível em:http://www.oaji.net/articles/2017/3932-1484338087.pdf

Cavalcante, A. P., & Cavalcante, M. M. P. (2010). *Tecnologias de disseminação da informação na web: um estudo sobre o Google Books. Múltiplos Olhares Em Ciência Da Informação - ISSN 2237-6658*, *3*(2). [Em linha] [Consult. 25 junho 2018] Disponível em: http://www.portaldeperiodicos.eci.ufmg.br/index.php/moci/article/view/2125

Clausner, C., Pletschacher, S., & Antonacopoulos, A. (2014). *Efficient ocr training data generation with aletheia. Proceedings of the International Association for Pattern Recognition* [Em linha] [Consult. 25 julho 2018] Disponível em: https://www.primaresearch.org/www/assets/papers/DAS2014_Clausner_OCRTrainingDataGeneration.pdf

Coneglian, C. S., & Santarem Segundo, J. E. (2016). *Europeana no Linked Open Data: conceitos de Web Semântica na dimensão aplicada das Humanidades Digitais*. *Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência Da Informação*, *22*(48), 88. [Em linha] [Consult. 25 junho 2018] Disponível em: https://doi.org/10.5007/1518-2924.2017v22n48p88

Crespo, L. B. (2001). *Bibliotecas digitales y actividad bibliotecaria. Ciencias de La Información, 32(1), 57–68.* [Em linha] [Consult. 15 abril 2018] Disponível em:http://cinfo.idict.cu/index.php/cinfo/article/viewArticle/255%5Cnhttp://cinfo.idict.cu/index.php/cinfo/article/viewFile/255/254

Demokritos. (2017). *The National Center for Scientific Research*. [Em linha] [Consult. 6 agosto 2018]. Disponível em: http://www.demokritos.gr/?lang=en

Europeana. (2018). Europeana Newspapers [Em linha] [Consult. 25 agosto 2018]. Disponível em: http://www.europeana-newspapers.eu/

Ferraz, M. N. (2014). *O papel social das bibliotecas públicas no século XXI e o caso da superintendência de bibliotecas públicas de Minas Gerais*. *Perspectivas Em Ciência Da Informação*, *19*(spe), 18–30. [Em linha] [Consult. 3 agosto 2018] Disponível em: https://doi.org/10.1590/1981-5344/2280

Giordano, R. B. (2016). *Do Jornal à Ciência: a Hemeroteca Digital Brasileira como fonte de informação para a pesquisa científica.* [Em linha] [Consult. 16 junho 2018] Disponível em:http://ridi.ibict.br/handle/123456789/883

Gorbea Portal, S. (2010). *Potencialidades de investigación y docencia iberoamericanas en ciencias bibliotecológica y de la información: memoria. Universidad Nacional Autónoma de México.* [Em linha] [Consult. 17 junho 2018] Disponível em:https://www.bibliotecologiaucr.wordpress.com/libros/

Herbert, J., & Estlund, K. (2008*). Creating citizen historians. Western Historical Quarterly*, *39*(3), 333-341. [Em linha] [Consutl. 30 julho 2018] Disponível em: https://scholarsbank.uoregon.edu/xmlui/handle/1794/9915

IFLA. (1998). *Principles for the Care and Handling of Library Material*. [Em linha] [Consult. 20 junho 2018]. Disponível em: https://www.ifla.org/files/assets/pac/ipi/ipi1-en.pdf .

IFLA. (2008). *Manifesto sobre Transparência, Bom Governo e Combate à Corrupção.* [Em linha] [Consult. 16 junho 2018]. Disponível em: https://www.ifla.org/files/assets/faife/publications/policy-documents/transparency-manifesto-pt.pdf

IFLA (2011) *Manifesto for Digital Libraries*. [Em linha] [Consult. 17 junho 2018] Disponível em: http://biblioo.info/wp-content/uploads/2012/11/Manifesto-IFLA.pdf

IFLA. (2014). *Guidelines for Planning the Digitization of Rare Book and Manuscript Collections.* [Em linha] [Consult. 17 junho 2018] Disponível em:

https://www.ifla.org/files/assets/rare-books-and-manuscripts/rbms-guidelines/guidelines-for-planning-digitization.pdf

IFLA. (1994) *Manifesto IFLA/UNESCO sobre Bibliotecas Públicas*. [Em linha] [Consult. 20 maio 2018] Disponível em: https://www.ifla.org/files/assets/public-libraries/publications/PL-manifesto/pl-manifesto-pt.pdfhttps://www.ifla.org/files/assets/public-libraries/publications/PL-manifesto/pl-manifesto-pt.pdf

IMPACT. (2018). *Centre of Competence*. [Em linha] [Consult. 2 agosto 2018]. Disponível em: https://www.digitisation.eu/.

International Association for Pattern Recognition. (1999). *Proceedings of the fifth International Conference on Document Analysis and Recognition.* [Em linha] [Consult. 6 agosto 2018]. Disponível em: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=791711

Lampoglia, F. (2012). *Discursividades da/sobre a ditadura militar em uma hemeroteca digital*. [Em linha] [Consult. 23 julho 2018]. Disponível em: https://repositorio.ufscar.br/handle/ufscar/1079

LIBER. (2018). Liga das Bibliotecas Europeias. [Em linha] [Consult. 30 agosto 2018]. Disponível em: https://www.libereurope.eu/liber-website-takedown-policy/

Library of Congress. (2011). *The National Digital Newspaper Program (NDNP) Technical Guidelines for Applicants.* [Em linha] [Consult. 26 julho 2018] Disponível em: http://www.loc.gov/ndnp/guidelines/archive/NDNP_201113TechNotes_rev2final.pdf

Library of Congress. (2018). *Ciclos anuais de prémios NDNP*. [Em linha] [Consult. 31 julho 2018]. Disponível em: http://www.loc.gov/ndnp/awards/

Library of Congress. (2018). *Chronicling America.* [Em linha] [Consult. 28 julho 2018]. Disponível em: https://chroniclingamerica.loc.gov/

Lima, I. F. de, Oliveira, H. P. C. de, & De Santana, S. R. (2013). *Metodologia para avaliação do nível de usabilidade de bibliotecas digitais: Um estudo na Biblioteca Virtual de Saúde*. *Transinformacao*, *25*(2), 135–143. [Em linha] [Consult. 3 agosto 2018] Disponível em: https://doi.org/10.1590/S0103-37862013000200004

Lima, I. F. de, Souza, R. R., & Dias, G. A. (2015). *Usabilidade da Biblioteca Virtual em Saúde: avaliando a eficácia, eficiência e satisfação. InCID: Revista de Ciência Da Informação e Documentação, 6*(1), 17. [Em linha] [Consult. 20 julho 2018] Disponível em: https://doi.org/10.11606/issn.2178-2075.v6i1p17-37

Mao, S. & Kanungo, T. (1999). *Empirical performance evaluation of page segmentation algorithms.* [Em linha] [Consult. 6 agosto 2018]. Disponível em: https://pdfs.semanticscholar.org/6668/1f964a53e5858053c08a7947fbf2765bad57.pdf

Marques, L. E., & Pinheiro, M. M. K. (2013). *A Cúpula Mundial sobre a sociedade da informação-CMSI: foco nas políticas de informação*. [Em linha] [Consult. 17 junho 2018] Disponível em:https://search.proquest.com/docview/1493870909?pq-origsite=gscholar

Martins, J. M. Q. (2008). Digitalização e guerra local: como fatores do equilíbrio internacional. [Em linha] [Consult. 20 julho 2018] Disponível em: https://www.lume.ufrgs.br/handle/10183/14405

Matos, A. (2000). *A digitalização do acervo documental da hemeroteca municipal de Lisboa: uma primeira abordagem ao suporte eletrônico, a partir do Jornal Os Ridículos. Colóquio Biblioteca e Novas tecnologias, Lisboa, Portugal, 7*. [Em linha] [Consult. 20 julho 2018] Disponível em: http://hemerotecadigital.cm-lisboa.pt/RecursosInformativos/ActasdeColoquiosConferencias/textos/digittexto.pdf

Medeiros, R., Melo, E. S. F., & Nascimento. (1990). *Trabalho oral: Impacto das Tecnologias de Informação na Gestão da Biblioteca Universitária. Uso estratégico das tecnologias em informação documentária. Hemeroteca Digital Temática: socialização da informação em cinema*. [Em linha] [Consult. 22 julho 2018] Disponível em: http://repositorio.ufrn.br:8080/jspui/bitstream/1/2964/1/SNBUHemerotecaCinema.pdf

Medeiros, R., Melo, E. S., & Nascimento, M. D. S. (2008). *Hemeroteca digital temática: socialização da informação em cinema.* [Em linha] [Consult. 22 julho 2018] Disponível em: http://www.repositorio.ufrn.br:8080/jspui/handle/1/2964

Milnovic, V., Trtovac, A., & Sofronijevic, A. (2014). *The Importance of the Digitized Serbian Periodicals in the Context of "Europeana Newspapers Project*". [Em linha] [Consult. 22 agosto 2018] Disponível em: http://itlit.net/ra1.pdf

Mühlberger, G. (1999). *Digitisation of newspaper clippings: The LAURIN project.*

Natinal Endowent for Humanities. (2018). [Em linha] [Consult. 28 julho 2018]. Disponível em: https://www.neh.gov/

Oliveira, J. B. de (2005). *Hemeroteca sobre saques e invasões: do impresso ao digital*. [Em linha] [Consult. 22 julho 2018] Disponível em: https://monografias.ufrn.br/jspui/bitstream/1/204/1/JulianaBO_Monografia.pdf

Papadopoulos, C., Pletschacher, S., Clausner, C., & Antonacopoulos, A. (2013). The IMPACT dataset of historical document images. *Proceedings of the 2nd International Workshop on Historical Document Imaging and Processing - HIP '13*, 123. [Em linha] [Consult. 14 agosto 2018] Disponível em: https://doi.org/10.1145/2501115.2501130

Patrício, H. S. (2010). *Desenvolvimento de serviços digitais na Biblioteca Nacional de Portugal: cinco perspetivas fundamentais*. [Em linha] [Consult. 10 agosto 2018] Disponível em: https://www.bad.pt/publicacoes/index.php/congressosbad/article/view/196/192

Pletschacher, S., & Antonacopoulos, A. (2010). The PAGE (Page Analysis and Ground-truth Elements) format framework. *Proceedings - International Conference on Pattern Recognition*, 257–260. [Em linha] [Consult. 14 agosto 2018] Disponível em: https://doi.org/10.1109/ICPR.2010.72

Pontes, F. V., & de Oliveira Lima, G. A. B. (2012). *A organização do conhecimento em ambientes digitais: aplicação da teoria da classificação facetada. Perspectivas Em Ciência Da Informação*, *17*(4), 18–40. [Em linha] [Consult. 25 julho 2018] Disponível em: https://doi.org/10.1590/S1413-99362012000400003

Rosetto, M. (2006). Bibliotecas digitais: cenários e perspectivas. *RBBD. Revista Brasileira de Biblioteconomia e Documentação*, *4*(1), 101–130. [Em linha] [Consult. 14 julho 2018] Disponível em: https://rbbd.febab.org.br/rbbd/article/view/101/92

Santos, P. dos. (2004). A dimensão política da Disseminação da Informação através do uso intensivo das tecnologias de Informação e Comunicação uma alternativa à noção de Impacto Tecnológico. *DataGramaZero - Revista de Ciência Da Informação*, *5*(4), 1–13. [Em linha] [Consult. 5 agosto 2018] Disponível em: http://dici.ibict.br/archive/00000352/01/A_dimenção_política_da_disseminação.pdf

Semonche, B. (2003). Newspaper Indexing Policies and Procedures, (October 27, 2008). [Em linha] [Consult. 1 agosto 2018] Disponível em:

http://parklibrary.jomc.unc.edu/indexing.html

Shaw, E. J., Blumson, S., & OCR, B. (1997). *Making of America-Online Searching and Page Presentation at the University of Michigan.* [Em linha] [Consult. 16 agosto 2018] Disponível em: http://www.dlib.org/dlib/july97/america/moadlib.html

The Usability Post. (2018*). Characteristics of Successful User Interfaces*. [Em linha] [Consult. 21 agosto 2018]. Disponível em: http://www.usabilitypost.com/2009/04/15/8-characteristics-of-successful-user-interfaces/

Toutain, L. B., Sayão, L., Marcondes, C., & Kuramoto, H. (2005). *Bibliotecas digitais: saberes e práticas.* [Em linha] [Consult. 17 julho 2018]. Disponível em: http://livroaberto.ibict.br/bitstream/1/1013/1/Bibliotecas%20Digitais.pdf

University of Salford. (2018). PRImA. [Em linha] [Consult. 25 agosto 2018]. Disponível em: http://www.primaresearch.org/tools