# A hemeroteca digital brasileira

**Angela Maria Monteiro Bettencourt** (BN) - crd@bn.br

**Monica Rizzo Soares Pinto** (BN) - rizzo@bn.br

**Resumo:**

*Esse relato de experiência descreve a problemática e a metodologia adotadas na criação da Hemeroteca Digital Brasileira, projeto financiado pela Financiadora de Estudos e Projetos (FINEP), com o objetivo de preservar e dar acesso a 10 milhões de páginas de periódicos brasileiros pertencentes à coleção da Biblioteca Nacional (BN). São descritas as diversas etapas da cadeia de desenvolvimento do projeto, entre elas a seleção, a captura, a descrição/representação, indexação, o uso de reconhecimento ótico de caracteres, encadernação virtual e o acesso aos arquivos digitais. São abordadas também a repercussão de seu lançamento, além das perspectivas para o futuro do projeto.*

**Palavras-chave:** *Curadoria digital. Acervos memoriais. Hemeroteca digital. Indexação conteúdos web. Acesso livre.*

**Área temática:** *Temática I: Tecnologias de informação e comunicação – um passo a frente*

# A hemeroteca digital brasileira

Resumo: Esse relato de experiência descreve a problemática e a metodologia adotadas na criação da Hemeroteca Digital Brasileira, projeto financiado pela Financiadora de Estudos e Projetos (FINEP), com o objetivo de preservar e dar acesso a 10 milhões de páginas de periódicos brasileiros pertencentes à coleção da Biblioteca Nacional (BN). São descritas as diversas etapas da cadeia de desenvolvimento do projeto, entre elas a seleção, a captura, a descrição/representação, indexação, o uso de reconhecimento ótico de caracteres, encadernação virtual e o acesso aos arquivos digitais. São abordadas também a repercussão de seu lançamento, além das perspectivas para o futuro do projeto.

Palavras chave: Curadoria digital. Acervos memoriais. Hemeroteca digital. Indexação conteúdos web. Acesso livre.

Área temática I: Tecnologias de informação e comunicação – um passo a frente

## 1 INTRODUÇÃO

Desde sua transferência de Portugal para o Brasil, no início do século XIX, a Biblioteca Nacional (BN), uma das mais importantes instituições de memória do país, coleta, preserva e franqueia o acesso à Memória Nacional. A coleção de periódicos da BN é a mais antiga, bem como, a mais completa do país. A instituição sempre contou com dispositivos legais para a formação de sua coleção. De acordo com Pinto (2011, p. 48-67), no Império, essa legislação, era restrita somente à Tipografia Nacional (1822), e, posteriormente foi estendida às tipografias sediadas na Corte (1847). Foi somente em 1907, que a legislação ganhou caráter nacional. A Lei do Depósito Legal em vigor determina o envio à BN de “todas as publicações, produzidas por qualquer meio ou processo, para distribuição gratuita ou venda” (BRASIL, 2004) editadas no Brasil ou com editor aqui domiciliado.

Destacam-se nessa coleção alguns jornais, tais como: *Correio da Manhã* (1901) - um dos mais importantes jornais da história da imprensa brasileira e o jornal extinto mais consultado na Biblioteca Nacional, *O Paiz* (1860) e a *Gazeta do Rio de Janeiro* (1808) - primeiro jornal publicado no Brasil. Dentre as revistas nacionais estão títulos que contribuíram para a formação da cultura e política brasileira, como:

a satírica *Careta* (1908); *O Malho* (1902) - a primeira revista brasileira a usar cor em suas páginas; *Revista da Semana* (1900) - a grande revista de variedades do início do século; *O Tico-Tico* (1905) - a primeira revista de histórias em quadrinhos, além de, *Illustração Brasileira* (1909). O acervo também é composto por periódicos de caráter científico, como: *Revista de Engenharia* (1879), *Vellosia* (1887), sobre botânica, *Diário de Saúde* (1835); *Semanário de Saúde Pública*: *pela sociedade de medicina do Rio de Janeiro* (1831); e *Revista dos Constructores: architectura e engenharia hygiene e pratica das construções* (1889), entre outros.

Como parte de sua política de preservação, a BN vem, desde meados da década de 1940, microfilmando seu acervo. Janice Monte-Mór (1977, p. 293-294) registrou que, a partir do início dos anos 1970, projetos procuraram salvaguardar coleções de periódicos pela microfilmagem continuada. Foram exemplos de sucesso a microfilmagem do *Jornal do Comércio* e dos Relatórios dos presidentes de províncias do período imperial. O Plano Nacional de Microfilmagem de Periódicos Brasileiros – PLANO, criado em 1978, e coordenado pela BN desde 1982, "tem como objetivo identificar, localizar, organizar, recuperar e preservar pela microfilmagem o acervo hemerográfico brasileiro existente nas diversas unidades da Federação visando sua recuperação para a Biblioteca Nacional, órgão depositário da memória impressa nacional, e facilitar-lhe a consulta" (ZAHER, 1983, p.315). Como consequência conseguiu, ao longo dos anos, reunir, resgatar e mapear grande parte da produção hemerográfica do país complementando virtualmente o acervo da Biblioteca Nacional e o tornando ímpar para a memória brasileira.

> "Um dos mais notáveis trabalhos da Biblioteca Nacional [...]. Articulando-se com outras bibliotecas do Rio de Janeiro e das diversas unidades da Federação, a Biblioteca Nacional recompôs, através da microfilmagem, coleções de periódicos raros, com seus exemplares muitas vezes dispersos em instituições diversas pelo país. O resultado é também um catálogo coletivo. Hoje são dezenas de milhares de microfilmes, com milhões de fotogramas, que registraram a história editorial brasileira do século XIX, preservando a informação e dinamizando o acesso, não só para os consulentes da sede da Biblioteca Nacional no Rio de Janeiro, mas atendendo pelo correio a leitores de todo o país e do exterior. Poucos países do mundo realizaram um projeto desse porte." (HERKENHOFF, 1996, p. 97)

Atualmente encontram-se microfilmados nove mil títulos de periódicos totalizando trinta e dois mil rolos de microfilmes, que equivalem a mais de trinta milhões de imagens.

## 2 METODOLOGIA

Com o patrocínio da FINEP, o projeto digitalizou e disponibilizou dez milhões de páginas de periódicos. Para sua execução foi desenvolvida uma metodologia própria contemplando as diversas etapas da cadeia criada para o desenvolvimento do projeto, entre elas a seleção, a captura, a descrição e representação, a indexação e a disponibilização dos arquivos digitais.

O projeto seguiu um cronograma de produção dividido em duas frentes, a primeira para periódicos em preto e branco onde a conversão para o digital se fez a partir do microfilme e a segunda para periódicos coloridos onde esta se deu a partir do documento original.

O tratamento dos arquivos digitais para fins de acesso contemplou o reconhecimento ótico dos caracteres dos conteúdos com a finalidade de potencializar a busca textual e refinar a recuperação da informação. O modelo de interoperabilidade seguido foi o mesmo já adotado pela BNDigital e baseado no protocolo da Iniciativa dos Arquivos Abertos (OAI-PMH), que permite a coleta e o intercâmbio de metadados entre repositórios digitais. Os padrões adotados garantem a preservação a longo prazo dos arquivos gerados.

### 2.1 Seleção

Os critérios de seleção basearam-se nos seguintes princípios: periódicos brasileiros, incluindo aqueles publicados fora do território nacional, como é o caso do *Correio Braziliense* (1808); periódicos em domínio público ou aqueles cujos direitos de publicação foram cedidos à BN como é o caso do *Jornal do Brasil*, *Conjuntura Econômica*, entre outros; periódicos raros e os periódicos mais solicitados pelos usuários para consulta local e para reprodução. A seleção procurou respeitar a legislação de direitos autorais.

### 2.2 Captura

Foram adotados dois critérios para a captura ou conversão para o digital, um para periódicos em preto e branco, e outro, para os coloridos. A captura dos

periódicos em preto e branco ficou a cargo da empresa DocPro e foi realizada na Biblioteca Nacional a partir do microfilme negativo que é o máster em microfilmagem. Foram utilizados escâneres de microfilmes FlexScan de alta produtividade, capazes de digitalizar vinte mil fotogramas por dia. A etapa de captura, executada nesses equipamentos, é quase completamente automatizada, prescindindo da interferência humana. Nas etapas posteriores de nomeação, separação e encadernação virtual, a atuação humana é fundamental, o que torna o trabalho mais lento e meticuloso. Se aumentada a produção diária de imagens digitais, as etapas seguintes tratamento e encadernação não se desenvolveriam em sincronia com a etapa de captura.

As características dos arquivos digitais máster gerados a partir de microfilme são:

a) resolução das imagens de 300 ppi[1], para um melhor reconhecimento ótico dos caracteres;
b) formato TIFF[2] para as imagens de guarda.

A digitalização dos periódicos coloridos foi feita a partir dos documentos originais no Laboratório de Digitalização da Biblioteca Nacional. Foram utilizados dois escâneres planetários de alta produção marca Zeutschel modelo 12000*HQ*. Esses equipamentos possuem diversas características que visam a salvaguarda do documento original. A média de captura diária de cada equipamento é de 500 imagens.

### 2.3 Nomeação

Visando a preservação e a disponibilização online, o projeto adotou o mesmo padrão para nomeação dos arquivos digitais adotado pela BNDigital: o identificador único do título do periódico, precedido pelo prefixo “per” que é a sigla convencionada pela BN identificadora do tipo de documento – no caso periódico + ano de publicação + número da edição + número sequencial da página. Essas regras valem tanto para os arquivos gerados a partir do microfilme, quanto para os arquivos gerados a partir de documentos originais.

### 2.4 Reconhecimento Ótico de Caracteres (OCR[3]) e encadernação virtual

[1] Pixels per inch
[2] Tagged Image File Format

Para a etapa de reconhecimento por OCR das 20.500 imagens geradas diariamente pelo projeto foram utilizadas 64 unidades de processamento funcionando 24 horas por dia. O software escolhido para o OCR foi o Abbyy FineReader 11 Professional.

A encadernação virtual foi feita fascículo por fascículo, espelhando o documento original, dessa forma o produto final é um arquivo PDF (*Portable document format*) pesquisável por edição.

## 2.5 Indexação e Recuperação da Informação

O projeto adotou dois processos de indexação e consequentemente duas formas de recuperação da informação. O primeiro processo é o já convencionalmente adotado pela BNDigital onde são indexadas as informações de: autoria, título, assuntos, datas e coleção. O segundo processo de indexação é novo na BNDigital e contempla o conteúdo do documento.

Para a indexação convencional o esquema de metadados utilizado para o projeto, assim como as normas e padrões adotados são os mesmos utilizados pela BNDigital. No quadro abaixo foram reunidos os metadados utilizados pela BNDigital e a norma ou o padrão adotado na descrição.

| BNDIGITAL | DUBLIN CORE | NORMA/PADRÃO ADOTADO |
|---|---|---|
| **Metadados de Identificação – Descritiva** | | |
| autoria principal e secundária - nome pessoal | <dc contributor> | AACR2 - Autoridades da BN |
| título principal | <dc title> | |
| formas variantes do título | <dc title> | |
| local | não atribuído | |
| publicador | <dc publisher> | Catálogo de Editores da BN |
| data | <dc date> | |
| descrição física | <dc description> | |

[3] Optical Character Recognition

| | | |
|---|---|---|
| Série | < dc source> | AACR2 - Autoridades da BN |
| nota geral | <dc description> | |
| nota de coleção | < dc source> | |
| tipo de suporte original | <dc type> | Norma BN |
| Idioma | <dc language> | Tabela – Languages |
| **Metadados de Identificação – Temática** | | |
| classificação decimal de Dewey | <dc subject> | CDD |
| assunto nome | <dc subject> | Assuntos da BN |
| assunto em inglês | <dc subject> | Assuntos da LC |
| **Metadados Administrativos** | | |
| nome arquivo digital | não atribuído | Norma BN |
| direitos | <dc rights> | Default: Biblioteca Nacional |
| formato arquivo | <dc:format> | MIME |
| escâner | não atribuído | Tabela BN |
| *software* | não atribuído | Tabela BN |
| compactação | não atribuído | Tabela BN |
| resolução | não atribuído | Default 300 ppi |
| tamanho em MB | não atribuído | |
| Cor | não atribuído | |
| intensidade de bits | não atribuído | |
| dimensões em pixels | não atribuído | |
| cópias segurança em HDs e DVDs | não atribuído | |
| Identificador | <dc identifiert> | URL |
| **Metadados Estruturais** | | |
| nota em | <dc source> | |
| nota de conteúdo | <dc relation> | |

Quadro1 - Representação da informação digital
Fonte: (BETTENCOURT, 2011, p.109)

Na indexação das palavras do conteúdo dos documentos foi utilizado o “Inteligenciamento DocPro”, processo que engloba a pesquisa por aproximação visual, característica principal da tecnologia DocPro, onde não são guardadas as palavras exatas e sim a aproximação visual de cada uma. Assim, as falhas que normalmente acontecem em um OCR comum são muito minimizadas, o que se traduz numa taxa de acerto em pesquisa muito superior.

## 2.6 Armazenagem e preservação

A armazenagem dos arquivos digitais, assim como sua preservação a longo prazo, são pontos cruciais para Bibliotecas Digitais de instituições que, como a Biblioteca Nacional, tem como missão a guarda da memória documental do país. Isso porque, atualmente, entende-se que esta missão compreende também a guarda dos documentos nascidos digitais. Da mesma forma que o impresso precisa de armazéns e estantes para sua armazenagem, o digital precisa de um centro de processamento de dados (*Data Center*) para ser armazenado de forma segura, permitindo o seu uso pelas futuras gerações.

O conceito de *Data Cente*r, hoje, se impõe como a solução mais adequada para armazenar dados. Formado por um conjunto de tecnologias que compõem uma estrutura responsável pelo armazenamento e também pelo processamento dos dados é um local projetado para ser extremamente seguro. Para que isso se concretize quatro fatores precisam trabalhar sincronizados:

a) infraestrutura com sistemas adequados e redundantes de climatização, energia, comunicação e monitoração;
b) segurança contra riscos físicos, que podem causar a paralisação das atividades, como incêndio, vazamento, jatos de água de combate, acesso indevido, roubo ou sabotagem;
c) manutenção preventiva e corretiva e de renovação do hardware realizadas por profissionais especializados e treinados;
d) certificação com as normas e regulamentações nacionais e internacionais de infraestrutura, segurança e manutenção.

A construção do centro de processamento de dados da BN - patrocinado pelo Banco Nacional de Desenvolvimento Econômico e Social (BNDES) - foi iniciada em 2011. A conclusão está prevista para o segundo semestre de 2013, quando passará a abrigar as bases de dados institucionais assim como os arquivos digitais produzidos pela BNDigital e por projetos de digitalização.

## 2.7 Disponibilização e acesso

O acesso à Hemeroteca Digital Brasileira pode ser feito através da BNDigital (http://bndigital.bn.br) ou através do site da Hemeroteca (http://hemerotecadigital.bn.br). Na BNDigital a busca pode ser por autor, título, editor e datas e palavras-chave. As edições ou fascículos são apresentados de forma cronológica através de interfaces semelhantes a calendários. Já a busca no site da Hemeroteca é novidade na BNDigital, pode ser feita por palavras-chave no conteúdo textual de uma determinada coleção, em um período de tempo específico ou em uma região ou cidade de publicação.


Figura 1 – Hemeroteca Digital Brasileira – página inicial

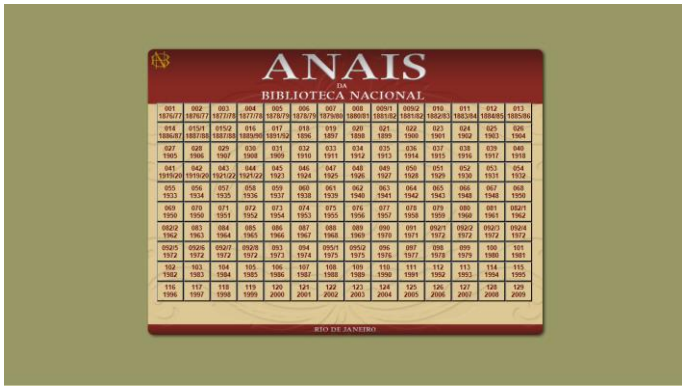


Figura 2 – Calendário

Outra novidade é o desenvolvimento de um dicionário de correspondências ortográficas que permitirá a recuperação de palavras escritas com grafias diferentes. Por exemplo, se digitarmos na busca a palavra “farmácia” o sistema recuperará também a palavra “pharmacia”.

O site da Hemeroteca também reúne artigos, escritos por especialistas contextualizando os principais periódicos digitalizados. O objetivo é ir além da simples disponibilização de fontes primárias de pesquisa, e apresentar também conteúdos inéditos contextualizando o acervo. Assim, pretende-se atender tanto o pesquisador tradicional, que tem um foco específico e já conhece o acervo digital da BN, quanto o internauta comum que pode chegar ao site através das ferramentas de busca na internet.

## 3. CONCLUSÃO

A Hemeroteca Digital Brasileira, parte integrante da Biblioteca Nacional Digital (BNDigital), materializa uma das tradicionais missões das Bibliotecas Nacionais: proporcionar o amplo acesso às informações contidas em seu acervo. Jornais, revistas, boletins, relatórios e outras publicações periódicas são fontes primárias de informação histórica – cultural, científica, técnica, política etc. –, trata-se, pois, de acervo de interesse público, que requer ampla difusão e fácil acesso por todos os cidadãos não só brasileiros como de todo o mundo. Os acessos mensais ao site da Hemeroteca, atualmente[4], giram em torno de 400 mil – número que, em tão pouco tempo, revela o alcance e o elevado significado dessa iniciativa.

O lançamento do projeto, em julho de 2012, com 5 milhões de páginas[5], repercutiu positivamente na mídia impressa e digital. A revista *Ciência Hoje* mencionou que “além da facilidade de acesso, a Hemeroteca Digital também traz um ganho econômico para quem precisa consultar os documentos históricos. Isso porque para tirar cópias do acervo físico são cobradas taxas, o que pode tornar uma pesquisa extensa bastante cara.”. *O Globo* destacou a abrangência do conjunto já que “costumes e mudanças no Rio da virada do século XIX para o XX são temas também facilmente explorados com a nova ferramenta.”

[4] Dado de abril de 2013
[5] Atualmente são cerca de 10 milhões de páginas

As perspectivas futuras incluem digitalizar mais 20 milhões de páginas, estabelecendo, parcerias com outras instituições públicas, a exemplo de parcerias de sucesso como, por exemplo, com o Instituto de Pesquisas Jardim Botânico do Rio de Janeiro[6].Implementar acordos com empresas e detentores de direitos autorais são também metas que visam ampliar exponencialmente acesso à pesquisa e ampliar a disponibilização de periódicos brasileiros extintos e correntes na Hemeroteca Digital Brasileira.

**Referências:**


BETTENCOURT, Ângela Maria Monteiro. **A representação da Informação na Biblioteca Nacional do Brasil: do documento tradicional ao digital.** Rio de Janeiro, 2011.

BRASIL. Lei n. 10.994, de 14 de dezembro de 2004. **Diário Oficial [da] República Federativa do Brasil**. Brasília, DF, 15 dez. 2004. Disponível em: <http://www.bn.br/bnPortal/site/rightView/LeidepositoLegal.htm>. Acesso em: 14 abr. 2013.

HUTFLESZ, Yuri. História virtual. **Ciência Hoje on Line**. Rio de Janeiro. Set. 2012. Disponível em: < http://cienciahoje.uol.com.br/revista-ch/sobrecultura/2012/09/historia-virtual>. Acesso em: 14 abr. 2013.

HERKENHOFF, Paulo. **Biblioteca Nacional**: a história de uma coleção. 2. ed. Rio de Janeiro: Salamandra, 1997. 263p.

LIMA, Ludmila. Uma coleção de raridades a alguns cliques. **O Globo**. 6 out. 2012. Disponível em: < http://oglobo.globo.com/rio/uma-colecao-de-raridades-alguns-cliques-6311864> Acesso em: 14 abr. 2013.

MONTE-MOR, Janice. Relatório, 1977. **Anais da Biblioteca Nacional**, v. 97, p.284-198. 1977. Disponível em: <http://objdigital.bn.br/acervo_digital/anais/anais_097_1977.pdf> Acesso em: 14 abr. 2013.

PINTO, Mônica Rizzo Soares. **Preservação de publicações eletrônicas: a questão do depósito legal.** Rio de Janeiro, 2011.

BIBLIOTECA NACIONAL (BRASIL). **Relatório técnico: convênio 0.1.10.0540.00**. Rio de Janeiro, 2012. 22 p.

ZAHER, Célia Ribeiro. Relatório, 1983. **Anais da Biblioteca Nacional**, v. 103, p. 307-335. 1983. Disponível em: <http://objdigital.bn.br/acervo_digital/anais/anais_103_1983.pdf> Acesso em: 14 abr. 2013.


[6] Essa parceria permitiu complementar o conjunto da revista *Rodriguesia*, disponibilizando a coleção completa do periódico na Hemeroteca