A gestão de obras digitalizadas na BND

José Luís Borbinha, Gilberto Pedrosa, João Penas, João Gil

National Library of Portugal
Jose.Borbinha@bn.pt; gfsp@ext.bn.pt; penas@ext.bn.pt; jgil@ext.bn.pt

Palavras chave: *XML, METS, XHTML, Digitalização, Bibliotecas Digitais, Preservação Digital.*

Resumo. Este artigo aborda o problema da gestão de obras digitalizadas na iniciativa Biblioteca Nacional Digital. O desenvolvimento de colecções de obras digitais e digitalizadas levanta problemas especiais ao nível da gestão de cópias digitalizadas de originais manuscritos ou impressos, que se pretendem abordar de forma a permitir flexibilidade, reutilização e longevidade. Para responder a este problema a BN decidiu adoptar um esquema XML designado de METS, tendo desenvolvido para o efeito um conjunto de ferramentas adequadas aos principais casos de uso associados, sendo aqui descritas duas delas, denominadas de PAPAIA e ContentE, assim como a interoperabilidade entre ambas.

Introdução

Este artigo aborda o problema da gestão de obras digitalizadas na BN – Biblioteca Nacional, tal como está sendo abordado na perspectiva da iniciativa BND – Biblioteca Nacional Digital [1]. No contexto da BND tem vindo a ser desenvolvido um trabalho de digitalização de manuscritos, obras impressas e de outras obras e materiais em suporte físico. Além disso tem-se procurado desenvolver uma política de depósito de cópias digitais de obras relevantes para a missão da BN de biblioteca patrimonial, compreendendo tanto as obras nascidas e criadas para publicação digital como as cópias digitais de obras destinadas a ser impressas.

O desenvolvimento de colecções de obras digitais e digitalizadas levanta problemas especiais ao nível dos processos da descrição e catalogação, os quais não são neste momento abordados. Na BND seguem-se para este fim as regras e procedimentos normais definidos para a BN e recomendados pela PORBASE [6], os quais se têm mostrado suficientes no geral. Os novos desafios aqui abordados localizam-se a montante desses problemas, nos casos em que estamos perante a criação de cópias digitalizadas de originais manuscritos ou impressos.

De um modo simples o problema será o de como organizar as imagens e os respectivos metadados que são produzidas num projecto de digitalização, isso de forma estruturada e que permita a reutilização fácil e flexível de toda esses conteúdos em qualquer contexto (tal como a criação de cópias em XHTML).

METS

A manutenção de uma biblioteca de objectos digitais exige a manutenção dos metadados sobre esses objectos. Os metadados necessários para utilizar e gerir com sucesso objectos digitais são diferentes e mais vastos que os metadados utilizados para gerir colecções de obras impressas e outros materiais físicos. Embora uma biblioteca possa manter metadados descritivos sobre um livro da sua colecção, o livro não se dissolverá numa série de páginas soltas caso a biblioteca não registar metadados estruturais sobre a organização do livro, nem os investigadores serão incapazes de avaliar o valor do livro se a biblioteca não anotar que o livro foi produzido numa dada imprensa.

O mesmo não pode ser dito para uma versão digitalizada do mesmo livro. Sem metadados estruturais, os ficheiros com imagens ou texto serão de pouca utilidade, e sem metadados técnicos sobre o processo de digitalização, os investigadores poderão ter dúvidas sobre a exactidão da reflexão do original que a versão digital oferece. Por questões de gestão interna, a biblioteca deve ter ainda acesso a metadados técnicos apropriados para lhe permitir refrescar e migrar os dados, garantindo a durabilidade dos recursos.

Fig. 1. Exemplo da estrutura de uma obra em METS (criado pela aplicação ContentE).

Para responder a este problema a BN decidiu adoptar um esquema XML designado de METS – Metadata Encoding and Transmission Standard [4], de que se apresenta um exemplo de aplicação na Figura 1. Para geração e processamento desse formato de metadados estruturais foi desenvolvido um conjunto de ferramentas adequadas aos principais casos de uso associados, conforme adiante se descreve.

O formato METS foi definido pela DLF [2], tendo sete secções principais:

- **Cabeçalho**: O cabeçalho contém metadados descrevendo o documento METS em si, incluindo informação como o criador, editor, etc.
- **Metadados Descritivos**: Esta secção pode referenciar metadados externos ao documento METS (como um registo UNIMARC ou EAD acessíveis num servidor na Internet), conter metadados embebidos, ou ambos. Múltiplas instâncias de metadados, internas ou externas, podem ser incluídos nesta secção.
- **Metadados Administrativos**: Esta secção oferece informação sobre como os ficheiros foram criados e armazenados, direitos de propriedade intelectual, metadados sobre o objecto original a partir do qual o objecto digital foi derivado, etc. Tal como os metadados descritivos, os metadados administrativos podem ser tanto externos ao documento METS como codificados internamente.
- **Secção de Ficheiros**: Esta secção lista todos os ficheiros que compõem o objecto. Elementos <file> podem ser agrupados em elementos de grupos de ficheiros <fileGrp>, para permitir a subdivisão de ficheiros por versão do objecto.
- **Mapa Estrutural**: Este mapa é o coração do documento METS. Ele esboça uma estrutura hierárquica para o objecto, e liga os elementos dessa estrutura a ficheiros com conteúdos e metadados referentes a cada elemento.
- **Ligações Estruturais**: Esta secção permite registar referências entre nós na hierarquia esboçada no Mapa Estrutural. Esta secção tem um valor particular na utilização do METS para arquivar sítios da Internet.
- **Comportamento**: Estas secções podem ser usadas para associar comportamentos para os conteúdos dos objectos METS. Cada comportamento é composto de uma definição abstracta e ainda de uma referência para o módulo de código executável que o concretiza.

O esquema METS oferece uma norma útil para a troca de objectos digitais entre repositórios. Adicionalmente, o METS oferece a possibilidade de associar um objecto digital com comportamentos ou serviços. A discussão anterior descreve o esquema, mas uma examinação mais detalhada da sua documentação é necessária para compreender todo o alcance das suas capacidades.

No contexto do modelo de referência do OAIS – Open Archival Information System [5], e dependendo da sua utilização, um documento METS pode ser usado como Pacote de Informação de Submissão (SIP), um Pacote de Informação de Arquivo (AIP) ou um Pacote de Informação de Disseminação (DIP).

Todas as obras depositadas na BND são assim estruturadas segundo o esquema METS. As obras digitais criadas no exterior são sujeitas processos de transformação, mas já as obras digitalizadas criadas na BN (ou noutros locais, mas eventualmente utilizando a tecnologia disponibilizada para o efeito pela BN, acessível livremente), são estruturadas tendo logo de raiz o METS como formato nativo. Uma das ferramentas que a BN desenvolveu para esse efeito (estruturar obras digitalizadas em METS) é a aplicação ContentE, a qual pode ser utilizada de raiz, ou em complemento

a uma outra denominada de PAPAIA. Na Figura 1 apresenta-se um exemplo de um ficheiro criado dessa forma.

PAPAIA

A aplicação PAPAIA tem por objectivo o tratamento inicial das obras logo após a sua digitalização. As suas funcionalidades são essencialmente as seguintes:

- **Renomeação:** Renomeação automática de ficheiros de imagem de acordo com normas e parâmetros configuráveis, tais como cota, número de sequência, etc.
- **Edição de metadados TIFF:** Edição dos cabeçalhos TIFF [7] das imagens.
- **Estruturação:** Associação das imagens à descrição estrutural de uma obra.

O PAPAIA tem uma interface gráfica análoga à da navegação de directórios de um sistema de ficheiros. Além da estruturação, é possível associar a cada imagem uma ou mais palavras-chave, o que pode ser usado para criar índices de navegação. Toda esta informação é registada num ficheiro XML, como apresentado na Figura 2 (ficheiro este que pode ser reutilizado pelo ContentE).

São registados três tipos de elementos nesta descrição:

- **level** - nível ou nó na hierarquia estrutural de uma imagem ou agrupamento;
- **image** - associação de uma imagem a um nível da árvore;
- **keyword** - palavra-chave atribuída a uma determinada imagem.

```
<?xml version="1.0" encoding="UTF-8" ?>
- <work xmlns="http://www.bn.pt/PapaiaSchema/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instan
  xsi:schemaLocation="http://www.bn.pt/PapaiaSchema/ http://teresa.bn.pt/schemas/Papaia">
  - <application>
      <builder app="PAPAIA" date="2004:11:26:02:37:24" version="1.3" />
    </application>
  - <root title="Root">
    - <level title="Introducao">
      - <level title="L-64003-P_0007_1_t0.tif">
          <image src="L-64003-P_0007_1_t0.tif" alt="" />
        </level>
      - <level title="L-64003-P_0008_2_t0.tif">
          <image src="L-64003-P_0008_2_t0.tif" alt="" />
        </level>
      - <level title="L-64003-P_0009_3_t0.tif">
        - <image src="L-64003-P_0009_3_t0.tif" alt="">
            <keyword>D. Dinis</keyword>
          </image>
        </level>
      - <level title="L-64003-P_0010_4_t0.tif">
          <image src="L-64003-P_0010_4_t0.tif" alt="" />
        </level>
      - <level title="L-64003-P_0011_5_t0.tif">
          <image src="L-64003-P_0011_5_t0.tif" alt="" />
        </level>
      </level>
    - <level title="Capitulo 1">
      - <level title="L-64003-P_0012_6_t0.tif">
          <image src="L-64003-P_0012_6_t0.tif" alt="" />
        </level>
      - <level title="L-64003-P_0013_7_t0.tif">
          <image src="L-64003-P_0013_7_t0.tif" alt="" />
        </level>
```

Fig. 2. Exemplo da estrutura descritiva de uma obra exportada pela aplicação PAPAIA.

O elemento "keyword" pode incluir as coordenadas do rectângulo envolvente que localiza a palavra na imagem respectiva, o que pode expandir as opções de pesquisa e apresentação da obra (essa informação pode ser gerada por outra aplicação também desenvolvida na BND, denominada de KIWI).

ContentE

A aplicação ContentE tem como objectivo suportar o processo de construção de estruturas formais de obras digitalizadas. Um projecto é constituído por um ficheiro XML de abertura e uma directoria "master", na qual são inseridos os diferentes ficheiros que compõem a obra, possivelmente em múltiplos tipos MIME (TIFF, GIF, JPEG, ASCII, PDF, etc.). Podem-se importar descrições de estruturas das obras do PAPAIA ou criar novas descrições de raiz, e ainda importar outros metadados.

É possível assim a partir da cópia "master" gerar várias cópias da obra em formato XHTML, podendo estas apresentar diferentes estruturas e formas de visualizações. É também possível criar múltiplos índices de visualização de uma obra, segundo o tipo de exploração e acesso que se pretender oferecer. Pode-se exportar a obra em METS, sendo no entanto possível utilizar outros formatos equivalentes desde que devidamente definidos. Na Figura 3 apresenta-se a arquitectura da aplicação, com um exemplo da sua interface na Figura 4.

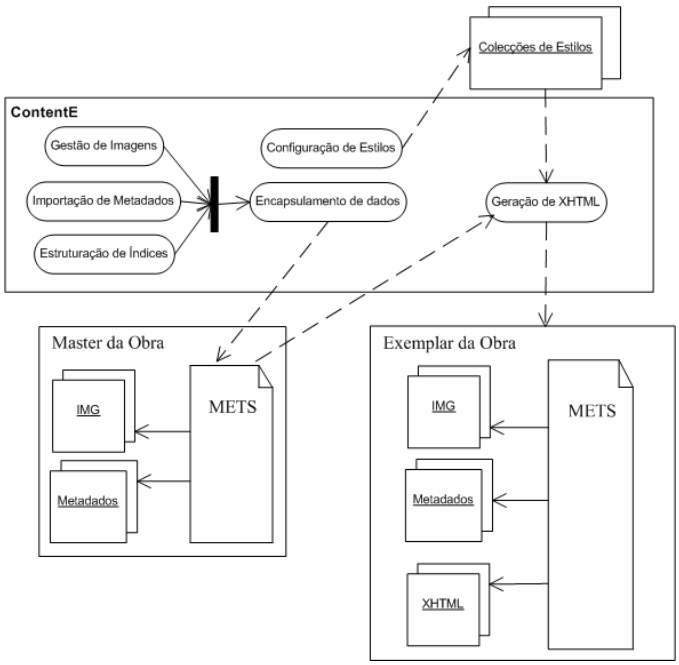

Fig. 3. Arquitectura da aplicação ContentE.

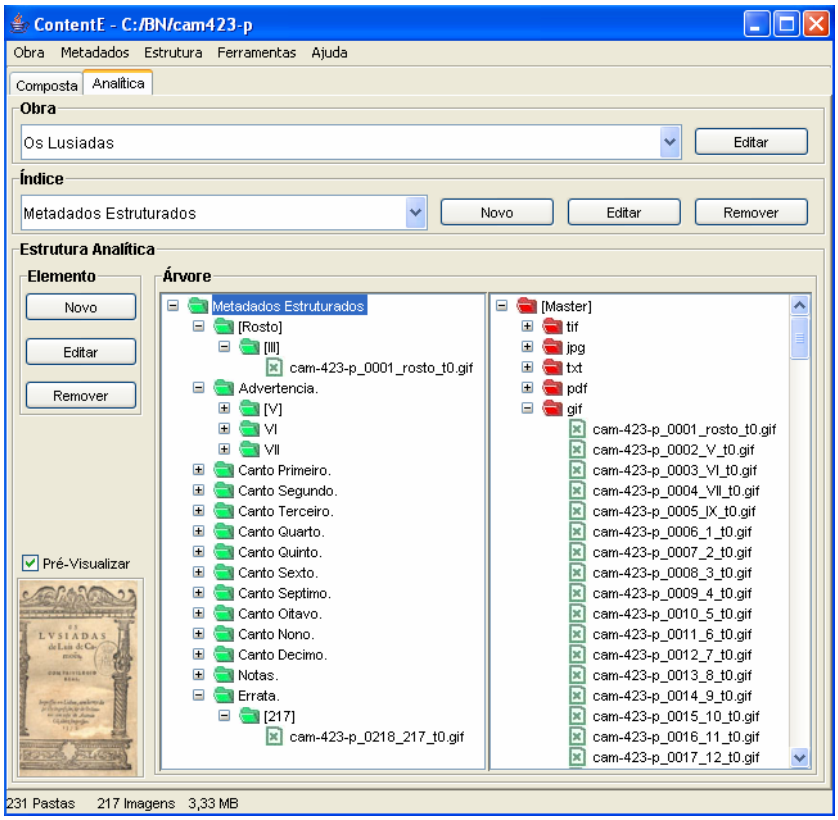

Fig. 4. Interface da aplicação ContentE.

Metadados

A cada obra digitalizada poderá ser associado um conjunto de diversos tipos de metadados. É possível associar diferentes metadados descritivos a diferentes nós. Os metadados importados poderão estar definidos em diferentes esquemas XML, sendo apenas necessário ter para cada caso o respectivo estilo XSL para a sua visualização. Exemplos desses metadados são metadados de direitos (a aplicar à obra no seu todo ou a partes específicas), técnicos (informação relativa ao processo de digitalização) ou descritivos (UNIMARC, Dublin Core, etc.).

No caso da BN, os metadados descritivos são importados em linha da PORBASE, através de um serviço web próprio, e associados à obra. Neste caso estes metadados são registados em MARCXML [3], como ilustrado na Figura 5.

Finalmente é possível registar ainda estruturas de metadados de termos e condições de uso das obras, os quais poderão ser aplicados a qualquer nó da estrutura (toda a obra, apenas um capítulo, apenas uma imagem, etc.). Neste momento a informação interpretada para este fim pela aplicação ContentE é registada no próprio METS, nos elementos definidos para o efeito, sendo no entanto possível registar

independentemente qualquer outra estrutura também (estas estruturas não são no entanto interpretadas, sendo utilizadas apenas para registo e visualização).

Fig. 5. Exemplo de um registo bibliográfico em MARCXML

Criação de cópias de visualização

No ContentE é possível definir os parâmetros de apresentação da obra, que são guardados em ficheiros XML seguindo um esquema criado para o efeito. A partir destas predefinições é possível criar várias configurações de apresentação (colecções de estilos) que podem ser modificadas e reutilizadas posteriormente, e que se adaptam a cada tipo de obra a gerar. Assim, torna-se necessário escolher o estilo de apresentação da obra antes da sua geração.

Antes da geração do novo exemplar é também produzido um ficheiro METS respectivo. Este ficheiro irá incluir apenas os conteúdos escolhidos especificamente para o exemplar a ser gerado e será inserido no directório do exemplar respectivo.

O passo seguinte consiste na aplicação de um estilo definido em XSL a este ficheiro METS. Este processo é repetido para cada página da obra, o que resulta na geração de um ficheiro XHTML por página. Para além destas, é ainda produzido uma página XHTML com a capa da obra e com os metadados descritivos. Também aqui, poderão surgir diferentes visualizações da informação recorrendo ao uso diferentes folhas de estilo, como ilustrado na Figura 6 e na Figura 7.

Os exemplares assim gerados podem ser manuseados de forma independentes, podendo ser transportados e consultados em qualquer lado.

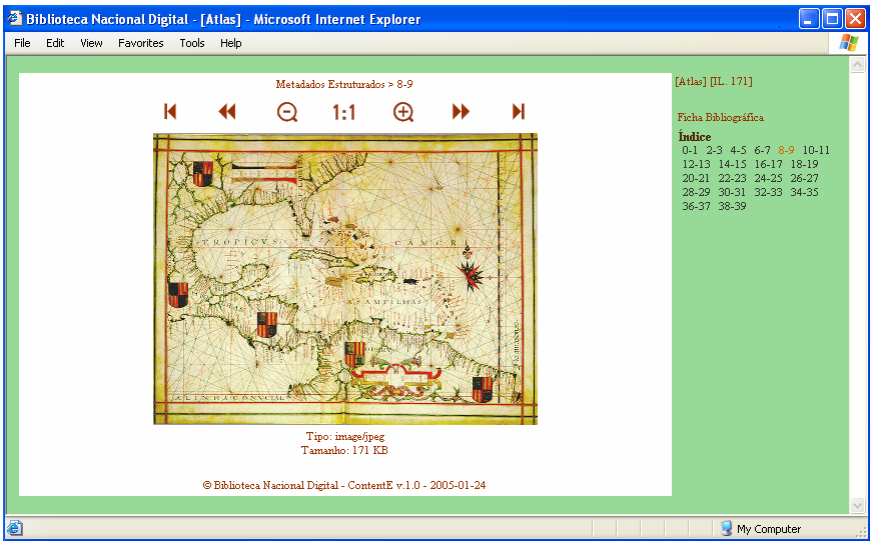

Fig. 6. Exemplo de uma obra publicada em formato XHTML

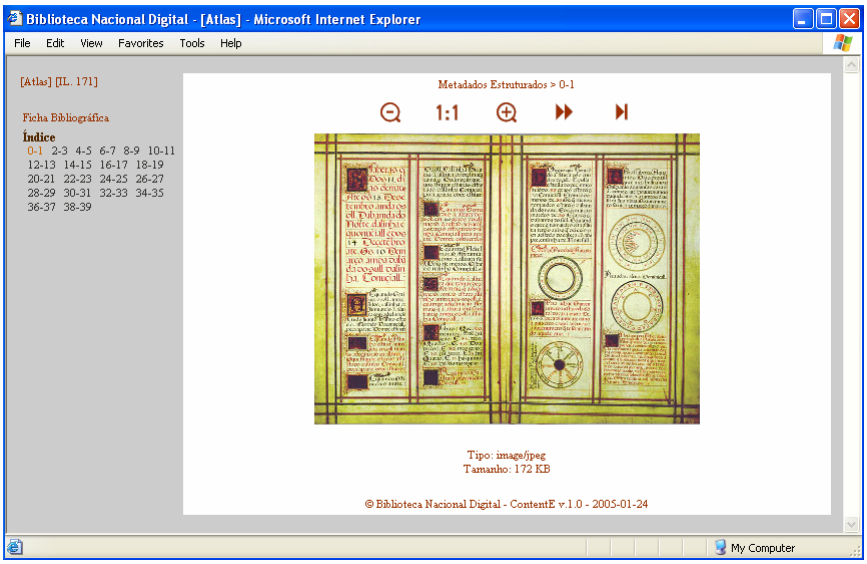

Fig. 7. Exemplo da mesma obra apresentada na Figura 6, mas agora com um outro estilo.

Conclusões

Os resultados aqui descritos correspondem a uma parte dos desenvolvimentos da segunda fase da iniciativa da Biblioteca Nacional Digital, que decorreu de meados de 2003 até ao final de 2004. Para além das provas de conceito terem demonstrado a validade dos modelos, foram ao mesmo tempo nesta fase desenvolvidas ferramentas e definidos processos igualmente válidos, suportando realmente os processos em curso.

Como trabalho futuro, fruto da aprendizagem entretanto adquirida e da reanálise dos processos de produção, foi já decidido integrar na aplicação Contente as funções da aplicação PAPAIA, visando obter um ambiente mais poderoso (as duas aplicações nasceram em contextos diferentes, tendo a sua convergência ocorrido naturalmente mas numa fase posterior às primeiras fases de análise).

Está planeado vir ainda a acrescentar ao ambiente ContentE a capacidade de importação e processamento de mais esquemas de metadados descritivos e estruturais, especialmente de descrições EAD [9] e TEI [10], assim como a capacidade de produção de local de cópias para acesso em PDF ou DAISY [11] (neste último caso associando ainda a capacidade de indexação de conteúdos sonoros).

Referências

1. BND – Biblioteca Nacional Digital [Em linha]. URL: <http://bnd.bn.pt>
2. DLF – Digital Library Federation [Em linha]. URL: http://www.diglib.org/dlfhomepage.htm
3. MARCXML – MARC21 XML Schema [Em linha] URL: http://www.loc.gov/standards/marcxml/
4. METS – Metadata Encoding and Transmission Standard [Em linha]. URL: http://www.loc.gov/standards/mets/
6. OAIS – Reference Model for an Open Archival Information System. [Em linha]. URL: http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html
7. PORBASE – Base Nacional de Dados Bibliográficos [Em linha]. URL: http://www.porbase.org
8. TIFF – Tag Image File Format reference v.6 [Em linha]. URL: http://partners.adobe.com/asn/developer/PDFS/TN/TIFF6.pdf
9. EAD - Encoded Archival Description [Em linha] URL: http://www.loc.gov/ead/
10. TEI - Text Encoding Initiative [Em linha] URL: http://www.tei-c.org/
11. DAISY Consortium [Em linha] URL: http://www.daisy.org/