Disponível em www.bad.pt/publicacoes

PAPER

12º CONGRESSO NACIONAL BAD

Evolução do modelo de informação da Biblioteca Nacional Digital

Helena Simões Patrício

Biblioteca Nacional de Portugal, hpatricio@bnportugal.pt

Resumo

Esta comunicação apresenta os aspetos mais significativos da evolução recente da estrutura de dados da Biblioteca Nacional Digital (BND), que se consubstanciam na definição de um novo modelo de objeto digital alinhado com o modelo de referência OAIS (Open Archival Information System); na aplicação de novos esquemas de metadados internacionais como o PREMIS (Preservation Metadata Implementation Strategies) e o MIX (Metadata for Images in XML); na formalização de um perfil de metadados para a BND; no desenvolvimento de uma ferramenta criada para a implementação deste novo modelo de informação digital e das operações de ingestão, acesso e administração do arquivo e, bem assim, na integração desse *software* com outas aplicações e sistemas de produção e gestão dos objetos digitais que constituem a BND.

Palavras-chave: Modelo de dados, Perfil de metadados, Modelo de referência OAIS, Preservation Metadata Implementation Strategies - PREMIS, Metadata Encoding and Transmission Standard - METS.

Introdução

O modelo de informação definido para a Biblioteca Nacional Digital (BND) compreende tanto o modelo de dados, que descreve a tipologia e os componentes dos objetos digitais, como o perfil de metadados, i.e., o registo das opções de implementação de um conjunto de diferentes esquemas, nomeadamente no que se refere aos elementos, atributos, valores, vocabulários controlados, restrições e articulação entre as diferentes entidades.

Na definição do modelo de informação da BND implementou-se o modelo de referência OAIS (Open Archival Information System), tendo-se ainda procedido à definição da tipologia e estrutura dos objetos digitais em alinhamento com a terminologia e conceitos subjacentes aos esquemas PREMIS (Preservation Metadata Implementation Strategies) e METS (Metadata Encoding and Transmission Standard). Por este motivo, começaremos por efetuar uma breve introdução a estes modelos, que permita perceber os seus conceitos mais importantes e compreender melhor os efeitos da sua aplicação na formalização de uma estrutura de dados para a BND.

Na sequência das alterações ao modelo de informação, a segunda parte da comunicação reporta o trabalho de formalização de um perfil de metadados para a BND, necessário para a definição das regras de integração dos diferentes esquemas de metadados, de forma a satisfazer os requisitos definidos para o modelo de dados e modelo funcional da BND.

A última parte do artigo refere-se ao processo de implementação do novo modelo de dados e do perfil de metadados, através do desenvolvimento da aplicação DZBTool, ferramenta que, para além de gerir os fluxos de trabalho de digitalização, publicação e armazenamento dos objetos digitais, assegura a produção dos próprios objetos digitais de acordo com a nova estrutura de dados.

Enquadramento conceptual

Para uma melhor contextualização teórica da nova estrutura de dados da BND, descreve-se em seguida o respetivo enquadramento no modelo de referência OAIS e, bem assim, os principais conceitos dos esquemas METS e PREMIS, utilizados na aplicação concreta daquele modelo aos objetos de informação da BND.

Modelo de referência OAIS

O modelo OAIS descreve a um nível abstrato os requisitos de um sistema de arquivo para preservação a longo prazo. Trata-se de um modelo funcional que descreve as operações, serviços e interfaces de seis áreas funcionais (ingestão, armazenamento de arquivo, gestão de dados, administração, planeamento de preservação e acesso) de um arquivo de preservação; mas é também um modelo de informação digital, que oferece os conceitos e a terminologia necessários à descrição e caracterização dos objetos de informação geridos pelo sistema.

O modelo OAIS utiliza o conceito de pacote de informação, para agrupar tanto os conteúdos de dados (objeto digital), como os metadados utilizados para descrever, preservar e aceder a esses dados. O pacote de informação coincide, assim, com o conceito de objeto de informação BND e inclui tanto o objeto de dados (conteúdos) como o objeto de metadados (descreve o conteúdo).

Cada objeto ou pacote de informação é, assim, constituído por informação de conteúdo (CDO, Content Digital Object) e por informação de metadados necessária à preservação a longo prazo dos objetos de informação: informação de representação (IR, Information Representation), informação de preservação (PDI, Preservation Description Information), informação de empacotamento (PI, Packaging Information) e informação descritiva (DI, Descriptive Information).

A informação de representação (IR) propicia a informação necessária para interpretar e apresentar os objetos de conteúdo (CDO) de forma a que os mesmos possam ser acessíveis, traduzindo os bits dos ficheiros de conteúdo em informação compreensível por quem a utiliza. Os conteúdos (CDO) e a informação de representação (IR) formam a denominada informação de conteúdo (CI, Content Information), sendo o alvo principal da preservação digital. Contudo, para a preservação a longo prazo são ainda precisos metadados adicionais como a informação de preservação (PDI), de empacotamento (PI) e descritiva (DI). A informação de preservação (PDI) consiste na informação de referência (identificadores únicos da informação de conteúdo dentro e fora do sistema de arquivo, por exemplo PURL, identificadores do objeto digital como a cota ou número de controlo bibliográfico do recurso analógico, etc), proveniência (origem e custódia do objeto de informação, ações de preservação e seus resultados), contexto (relação do objeto com o seu ambiente e com outros objetos de informação, por exemplo identificação de versões em formatos alternativos) e integridade (mecanismos que garantem a documentação de qualquer alteração ao objeto, por exemplo *checksum* dos ficheiros de imagem). A informação de empacotamento (PI) agrupa a informação de conteúdo e de metadados num único pacote de informação, documentando o relacionamento entre os dois tipos de ficheiros. A informação descritiva (DI) não integra os pacotes de informação, antes os descreve, de modo a que os mesmos possam ser descobertos pelas ferramentas de pesquisa e recuperação de informação pelos utilizadores (por exemplo, metadados descritivos contidos nos ficheiros *unimarc.xml*)

A tipologia dos pacotes de informação OAIS está relacionada com a função que os mesmos desempenham no arquivo digital: pacotes de submissão (SIP, Submission Information Package), pacotes de arquivo (AIP, Archival Information Package) e pacotes de disseminação (DIP, Dissemination Information Package).

Um pacote de informação de submissão SIP consiste na informação de conteúdo (CI) entregue pelo produtor para ingestão no arquivo. No modelo de informação da BND estes pacotes SIP correspondem às imagens matrizes (JPEG 2000 ou TIFF) a partir das quais é produzido o objeto de informação pela ferramenta DZBTool.

Os pacotes de informação de arquivo (AIP), correspondem à informação armazenada e preservada no repositório. Cada pacote AIP tem de conter toda a informação de preservação (OAIS-PDI) para determinado conteúdo de informação (CI). Os pacotes AIP da BND são o item matriz e o item master.

Os pacotes de informação de disseminação (DIP) são recebidos pelo utilizador em resposta a um pedido de informação e podem ser derivados de um ou mais pacote de arquivo AIP. Um pacote DIP pode não ter nenhuma informação de preservação (PDI). Na BND são pacotes DIP todos os itens de consulta e, para os utilizadores que operam no sistema OAIS, o item matriz.

Enquadramento METS

O esquema METS foi criado em 2001 pela Digital Library Federation, sendo atualmente mantido pela Biblioteca do Congresso. Trata-se de um standard *de facto* para o empacotamento e estruturação de objetos digitais, registado na NISO desde 2004 e que implementa o modelo OAIS, em harmonia com o qual foi, de raiz, desenhado. Possibilitando a construção de contentores XML capazes de compreender total ou parcialmente a estrutura de dados de um objeto digital, o METS é utilizado na BND para descrever tanto o objeto de informação na sua totalidade (ficheiro *metsItems.xml*), como cada item que o constitui (ficheiros *mets.xml*). Para além desta estruturação das partes componentes de um objeto, o METS é ainda utilizado para referenciar todos os ficheiros de conteúdo e de metadados que compõem esse objeto.

Um documento METS é constituído por seis secções: o *Header*, que contém metadados administrativos sobre a criação do próprio documento METS (agente, data, etc); a secção *dmdSec*, que inclui os metadados descritivos; a *amdSEC*, que engloba os metadados técnicos (techMD), de direitos (rightsMD), de recursos analógicos (sourceMD) e de proveniência digital (digiprovMD); a secção *fileSec*, que compreende todos os ficheiros de conteúdo do objeto, ordenados em grupos; a secção *structMap*, que especifica a estrutura do objeto e a localização dos ficheiros nessa estrutura; a secção *behaviorSEC*, que elenca todos os comportamentos de disseminação. A única seção obrigatória num documento METS é a *structMap*, tendo no modelo de informação da BND sido implementadas as seguintes secções: *Header*; secção *dmdSEC*, que referencia os metadados descritivos da obra digitalizada; divisão techMD, que localiza os metadados técnicos do objeto ou pacote de informação a que o ficheiro mets se aplica; divisão righstMD, relativa aos direitos de acesso aos diferentes itens do objeto; *seção fileSEC*, que elenca todos os ficheiros de conteúdo (estrutura física) e todos os itens do objeto (estrutura lógica); e, por último, a secção *strucMAP*, que localiza todos os elementos METS atrás descritos, dentro do próprio ficheiro de metadados.

Modelo conceptual PREMIS

Para a implementação do modelo de informação digital da BND adotou-se o esquema PREMIS, enquanto modelo semântico para metadados de preservação que não está vinculado a nenhuma estratégia específica de preservação digital.

O modelo PREMIS (*Preservation Metadata: Implementation Strategies*) foi criado em 2005 por uma equipa de trabalho conjunta da OCLC (Online Computer Library Center) e do RLG (Research Library Group), sendo atualmente mantido pela Biblioteca do Congresso.

Explicitamente enquadrado no modelo OAIS, o PREMIS especifica um conjunto de elementos de metadados que permitem implementar aquele modelo de referência. Esses elementos organizam-se em cinco entidades fundamentais: Entidades Intelectuais, Objetos, Eventos, Direitos e Agentes. Considera-se Entidade Intelectual um conjunto de conteúdos que pode ser considerado unitariamente para efeitos de descrição e gestão, por exemplo, um livro, um mapa, etc. Uma Entidade Intelectual pode ter uma ou mais representações digitais e um Objeto é uma unidade discreta de informação em formato digital. Entende-se por Evento, qualquer ação que envolva ou tenha impacto sobre um Objeto ou Evento contido ou associado ao repositório de preservação. Os Agentes são as pessoas, organizações ou programas de *software* associados aos Eventos ou aos Direitos associados ao Objeto e os Direitos são as permissões pertencentes a determinado Objeto e/ou Agente (PREMIS, 2012).

No modelo de informação BND, consideramos como Entidade Intelectual o recurso analógico digitalizado; na versão 2.3 do PREMIS implementada na BND não era possível representar esta entidade, pelo que a mesma é descrita não com recurso ao PREMIS, mas antes aplicando o esquema MarcXchange.

A entidade Objeto pode ser de três tipos: um ficheiro ou sequência ordenada de bytes reconhecida por um sistema operativo; um conjunto de bits dentro de um ficheiro; um conjunto de ficheiros que representam a entidade intelectual e que incluem metadados estruturais (PREMIS, 2012). No modelo de informação da BND apenas são consideradas as duas primeiras categorias de objetos PREMIS: ficheiros isolados e representações, uma vez que o objeto conjuntos de bits tem um nível demasiado granular de informação.

Os Eventos são entidades que agregam metadados sobre ações que modificam o objeto digital, por exemplo a produção de uma nova versão digital; ações que criam novas relações ou alteram relacionamentos já existentes; ações que não consubstanciam nenhuma alteração mas que verificam ou validam a integridade dos objetos (PREMIS, 2012). Não estando neste momento implementadas na BND medidas de preservação digital posteriores à ingestão dos objetos de informação no repositório, não foram por enquanto previstos elementos das entidades Eventos ou Agentes.

Os Direitos PREMIS codificam informação que permite a realização de determinada ação de preservação, diferindo, assim, dos metadados de direitos constantes de outros esquemas que se destinem a registar informação associada a ações de acesso/utilização dos objetos por parte de utilizadores externos ao sistema ou de operadores do repositório (PREMIS, 2012).

Modelo de dados BND

Apresentado o enquadramento teórico que fundamentou a conceção de um novo modelo de informação da BND, analisamos agora a respetiva implementação ao nível da estrutura ou modelo de dados, abordando na secção seguinte o perfil de metadados definido para a BND.

Os objetos de informação da BND têm uma natureza complexa, sendo cada objeto constituído por várias representações digitais da mesma obra. Estas representações constituem-se como sub-objetos ou pacotes de informação, no sentido definido pelo modelo OAIS, e que no repositório da BND são denominados por itens: item matriz, item master e itens de consulta.

O item matriz e o item master são pacotes de informação de arquivo (OAIS-AIP), uma vez que o destino principal dos respetivos conteúdos é a preservação digital. O item matriz e os itens de consulta são pacotes de informação de disseminação (OAIS-DIP), pois servem a finalidade de disponibilizar conteúdos tanto para acesso dos operadores do repositório, no primeiro caso; como para acesso do público, nos casos restantes.

O item matriz grupa as imagens matrizes do objeto e correspondentes metadados; o master contém as imagens que podem ser utilizadas para gerar as versões de consulta do objeto de informação e, bem assim, os metadados comuns a todos os itens; por ultimo, os itens de consulta contêm os elementos que permitem montar o pacote de disseminação dos conteúdos.

Cada um destes itens é um pacote de informação que engloba conteúdos, metadados e informação de empacotamento.

Os ficheiros de conteúdo correspondem ao material digital que se pretende preservar, consistindo em ficheiros de imagem e a ficheiros PDF. Estes conteúdos integram os pacotes de informação de arquivo (OAIS-AIP) da BND: o item matriz (agrupa as imagens matrizes, i.e. aquelas que garantem a preservação dos ficheiros resultantes da captura de imagem, de forma a não condicionar utilizações futuras, e que na BND assumem o formato TIFF ou JPEG 2000) e item master (contém as imagens de consulta derivadas das matrizes, que permitem constituição de pacotes de disseminação pelo público em geral). Os conteúdos estão agrupados em pastas por formato de ficheiro (cfr. ❶ na Figura 1).

Para além dos ficheiros de imagem, os objetos de informação são constituídos por metadados, que descrevem o conteúdo, o contexto e estrutura desses recursos, permitindo a interação com o objeto sem que haja conhecimento prévio da sua existência ou características (Foulonneau, 2008). Em regra, os metadados de um pacote de informação localizam-se na raiz do item (v. ❷ na Figura 1). Contudo, os metadados técnicos relativos ao objeto de dados (OAIS-CDO) e que correspondem à informação de representação (OAIS-IR) daqueles conteúdos localizam-se na pasta dos ficheiros de imagem/PDF a que respeitam (cfr. ❸ Figura 1).

Por último, a informação de empacotamento do objeto de informação corresponde ao ficheiro *metsItems.xml*, que referencia todos os items, ficheiros de imagem e de metadados do objeto, procedendo, assim, ao seu "empacotamento" global, de modo a que todas as partes componentes da sua estrutura estejam claramente identificadas e localizadas. Por este motivo, o ficheiro *metsItems.xml* não está localizado na raiz de nenhum item em concreto, mas sim na pasta geral das representações de consulta (cfr. ❹ Figura 1).

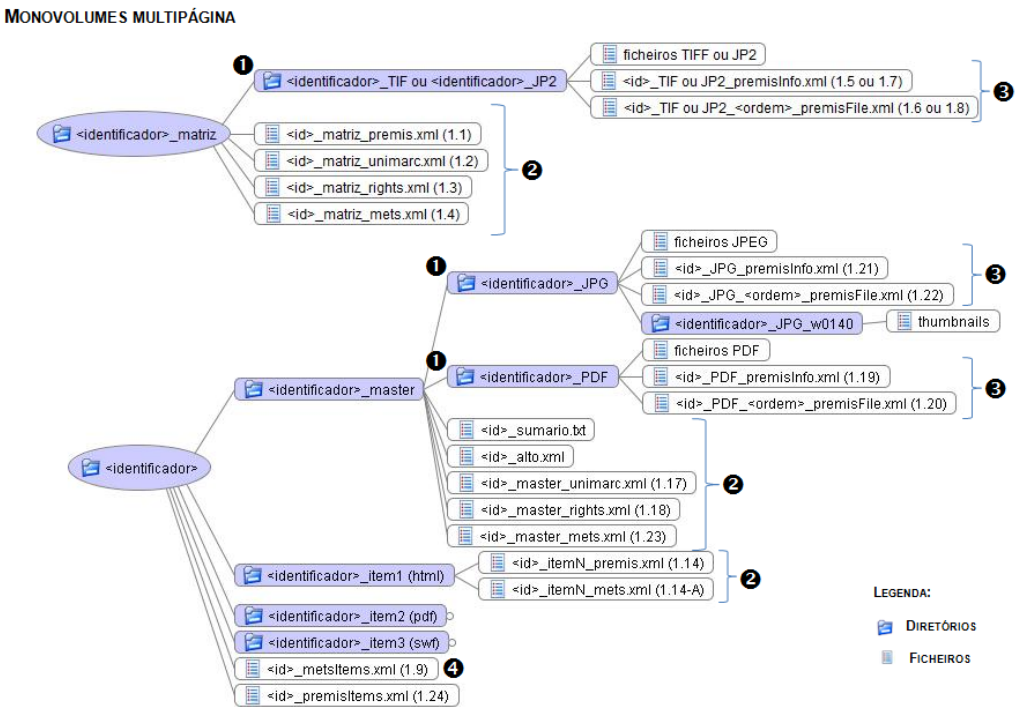

Figura 1 - Modelo de dados BND: exemplo de estrutura de obra multipágina

Uma descrição mais pormenorizada dos itens e ficheiros que compõem o modelo de dados BND, bem como a especificação da estrutura dos objetos de informação por tipologia, encontra-se documentada em *Modelo de informação da Biblioteca Nacional Digital*… (Patrício, 2015).

Face ao modelo de dados anteriormente usado na BND, aplicado a todos os objetos produzidos com a aplicação ContentE (Pedrosa, 2010), esta nova estrutura de dados apresenta as seguintes vantagens e diferenças fundamentais: simplificação da estrutura de diretórios e do nome dos ficheiros; maior rigor na identificação de ficheiros de metadados, que passaram a incluir a identificação do item a que correspondem, evitando-se assim a proliferação de ficheiros com o mesmo nome num mesmo objeto; eliminação de redundâncias na representação de metadados e imagens, e individualização de informação pertinente em ficheiros dedicados.

Com efeito, com o aumento generalizados da largura de banda, deixámos de ter necessidade de produzir imagens em mais do que uma resolução ou profundidade de cor, o que nos permitiu prescindir das subpastas dentro de cada formato de imagem e, bem assim, simplificar o nome dos ficheiros de imagem, eliminando aí os dados relativos às propriedades técnicas. Por outro lado, eliminaram-se pastas intermédias de metadados que não tinham utilidade, conforme pode observar-se na Figura 2.

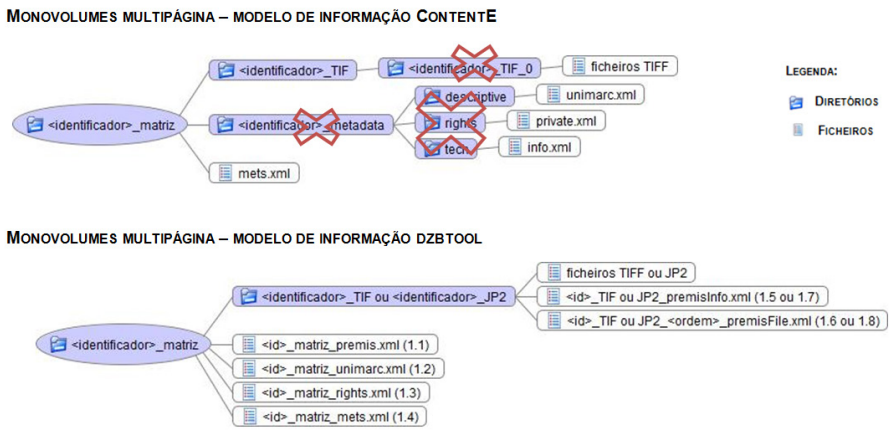

Figura 2 - Evolução do modelo de dados - Estrutura de diretórios

Um dos aspetos mais relevantes do novo modelo de dados é, contudo, o facto de ter deixado de haver duplicação de imagens e metadados no master e em todos os itens de consulta, como acontecia anteriormente. Efetivamente, as imagens e os metadados comuns a todos os itens de consulta passaram a estar representados apenas no item master.

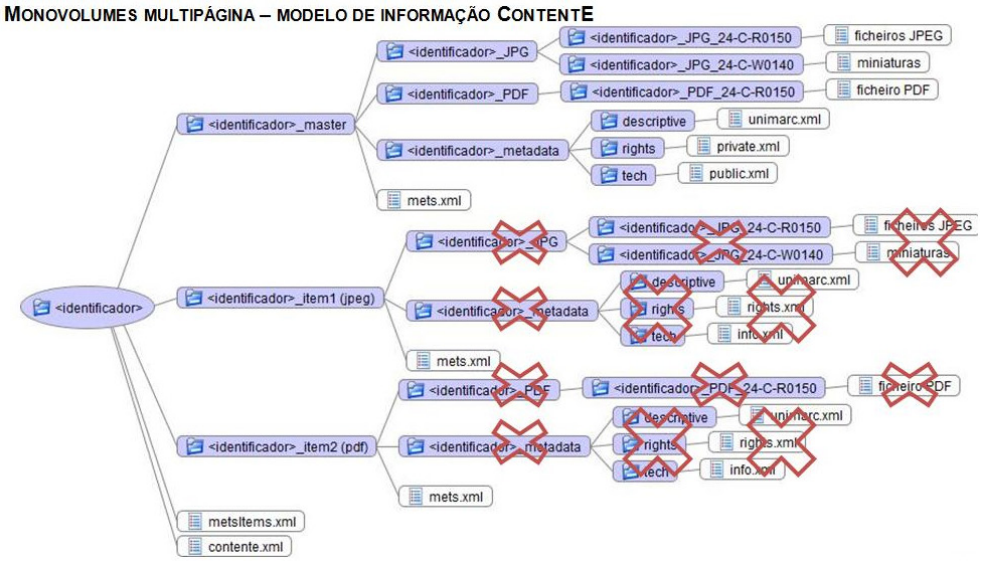

Figura 3 - Modelo de dados ContentE - Redundância de informação

Outro aspeto relevante no novo modelo de dados é incluir um ficheiro individual com os dados do «Sumário» ou estrutura lógica de um determinado item e, bem assim, do resultado do reconhecimento ótico de caracteres (OCR) num ficheiro codificado no *standard* aplicável a este tipo de informação, o esquema ALTO (ALTO, 2014). Este facto facilita a reutilização da estrutura do sumário e do texto integral do documento por outras aplicações ou pela própria ferramenta de produção do objeto.

Por outro lado, neste novo modelo prevê-se a existência de um ficheiro de metadados técnicos para cada imagem, ao contrário do que sucedia no modelo ContentE, em que a informação técnica das várias imagens está reunida num único ficheiro de metadados. Como veremos adiante a propósito da utilização do esquema MIX, os ficheiros de metadados técnicos relativos a imagens matrizes têm ainda a vantagem de colocar fora do ficheiro de imagem metadados que anteriormente estavam apenas embebidos nas etiquetas dos ficheiros TIFF.

Perfil de metadados da BND

Na definição de um perfil de metadados para a BND visámos documentar os esquemas e conjuntos de elementos escolhidos para a descrição dos objetos de informação, adicionando regras e orientações específicas para a utilização desses elementos, de modo a assegurar que as aplicações que criam e gerem esses objetos possam cumprir os seus requisitos funcionais.

Os diferentes esquemas de metadados selecionados para os objetos da BND podem ser agrupados de acordo com a seguinte tipologia:

- Esquema para o empacotamento de objetos de informação complexos e dos items que os constituem: esquema METS;
- Esquema para referenciação de objetos e itens a preservar: esquema PREMIS;
- Outros esquemas de representação: de informação técnica (PREMIS, MIX, TIFF tags, ALTO); de informação de direitos (METSRights, PREMIS); e de informação descritiva (MarcXchange).

A Tabela 1, abaixo, resume a correspondência entre os diferentes esquemas, os ficheiros de metadados que os implementam e o respetivo enquadramento no modelo OAIS.

Modelo BND	Tipologia metadados	OAIS	Esquema
❶ premisInfo.xml premiFile.xml ALTO.xml	Técnicos	Inf. Representação (IR)	TIFF tags; PREMIS; MIX Alto
❷ unimarc.xml	Descritivos	Inf. Descritiva (DI)	MarcXchange
❸ rights.xml	Direitos	Inf. Preservação (PDI)	METS Rights
❹ premis.xml premisItems.xml premisInfo.xml premiFile.xml	Preservação	Inf. Preservação (PDI)	PREMIS
❺ metsItems.xml mets.xml	Estruturais	Inf. Empacotamento (PI)	METS

Tabela 1 - Correspondência metadados BND, OAIS, esquemas

Analisando os metadados por tipologia, referimo-nos em primeiro lugar aos metadados técnicos, que permitem interpretar, apresentar e utilizar um recurso digital, e que descrevem os atributos dos conteúdos digitais ou imagens, tanto no que respeita ao processo de captura, como quanto ao ambiente técnico de processamento de conteúdos. Na BND, estes metadados estão contidos nos seguintes ficheiros (cfr. ❶ e ❹ na Figura 4).

Os metadados descritivos, que permitem descobrir e identificar o documento analógico que foi digitalizado, correspondem na BND aos ficheiros *unimarc.xml,* que guardam a informação constante do registo bibliográfico no momento da criação do objeto de informação. Cada objeto de informação contém dois ficheiros *unimarc.xml* exactamente com o mesmo conteúdo (cfr. ficheiros identificados com ❷ na Figura 4), sendo um armazenado no item matriz e outro no item master.

Os metadados de direitos permitem identificar o estatuto de direitos de autor do recurso ou permissões internas relativas à utilização de matrizes, para efeitos de implementação de restrições de acesso e utilização das suas cópias digitais, quer pelo público, no que se refere aos pacotes de disseminação, quer pelos operadores do sistema, no que se refere aos pacotes de arquivo. Cada objeto de informação contém dois ficheiros *rights.xml* diferentes (cfr. ❸ na Figura 4), sendo um armazenado no item matriz e outro no item master.

No que respeita aos metadados de preservação, que correspondem à informação de que um repositório necessita para operar os seus processos de preservação digital, cumpre referir que os mesmos se fundamentam sobretudo nos ficheiros *premisFile.xml* e *premisInfo.xml* referenciados anteriormente como metadados técnicos, e que estão especificamente contidos nos ficheiros *premis*.xml e *premisItems.xml* (cfr. ficheiros ❹ na Figura 4). Existe um ficheiro *premis.xml* por pacote de disseminação, i.e., existe um ficheiro *premis.xml* relativo ao item matriz e outro(s) por cada item de consulta. Em cada objeto BND existe, ainda, um ficheiro *premisItems.xml*, que identifica para preservação o ficheiro de metadados estruturais *metsItems.xml,* que empacota todo o objeto de informação digital.

Por último, temos os metadados de empacotamento ou estruturais, que descrevem a estrutura interna de um objeto, referenciando todas as suas partes componentes (items), identificando todas as tipologias de metadados de determinado objeto e, bem assim, todas as suas imagens ou conteúdos. Cada objeto de informação BND contém um ficheiro *mets.xml* para estruturação de cada item (matriz, master e consultas) e um único ficheiro *metsItems.xml* relativo ao objeto no seu todo (cfr. ficheiros ❺ da Figura 4).

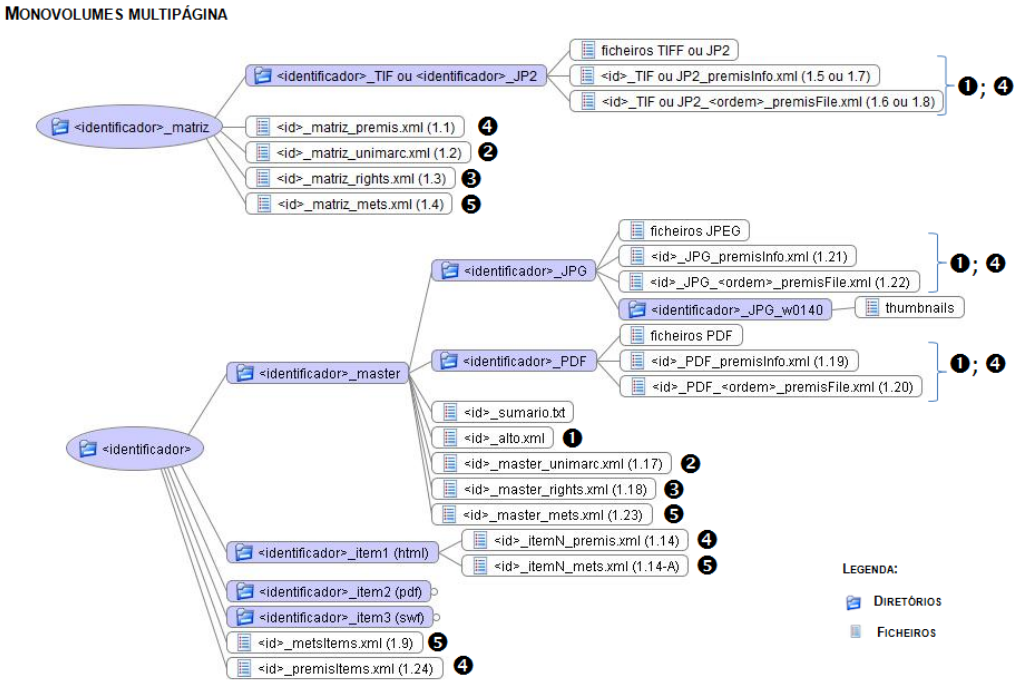

Figura 4 - Localização de metadados na estrutura de objeto BND

Implementação dos esquemas de metadados

Seguidamente, apresentaremos a implementação na BND dos esquemas de metadados METS, PREMIS e outros com eles relacionados.

Na criação de metadados estruturais aplica-se a versão 1.10 do esquema METS (http://www.loc.gov/standards/mets/mets.xsd), o *standard de facto* para o empacotamento de informação sobre objetos complexos, como é o caso na BND. Cada ficheiro METS representa um pacote de informação do objeto (*mets.xml*) ou o objeto na sua totalidade (*metsItems.xml*), permitindo identificar e localizar todas as suas partes componentes e a estrutura de ligações entre ficheiros de metadados e de conteúdos. Os ficheiros METS localizam os ficheiros de metadados descritivos (dmdSEC), de direitos (rightsMD) e técnicos (techMD), como se mostra na Figura 5.

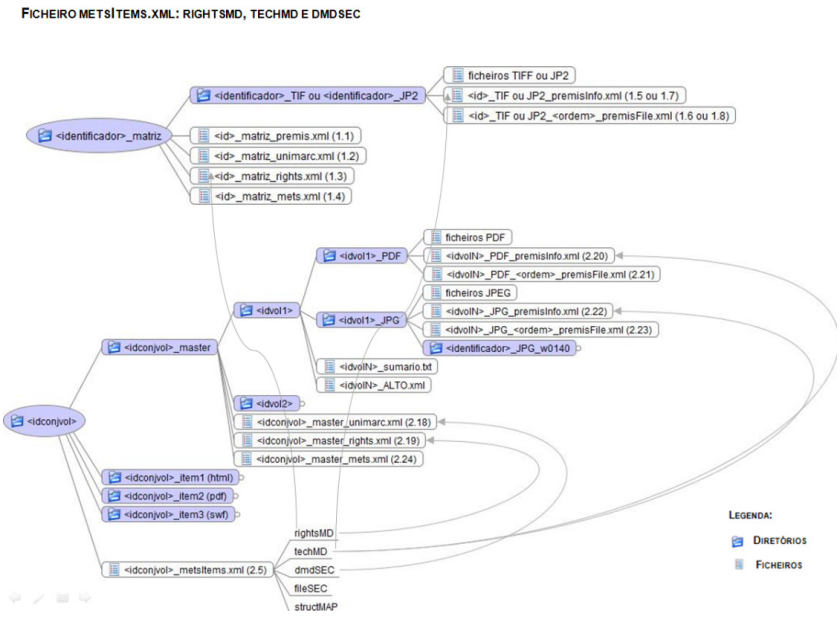

Figura 5 - Ficheiro metsItems: rightsMD, techMD e dmSEC

Através da secção *fileSEC* são localizados os ficheiros de conteúdo do objeto, tanto no que se refere aos ficheiros físicos de imagem (fileGRP:images); como quanto aos itens matriz, master e de consulta (fileGRP: itens) no caso do ficheiro *metsItems.xml*, sendo cada item representado pelo ficheiro *mets.xml* que empacota os diferentes ficheiros que o compõem. Por último, a secção s*trucMAP* do *metsItems.xml* permite localizar dentro do próprio ficheiro de metadados todas as "secções" e elementos identificativos de metadados, itens e imagens atrás referenciados.

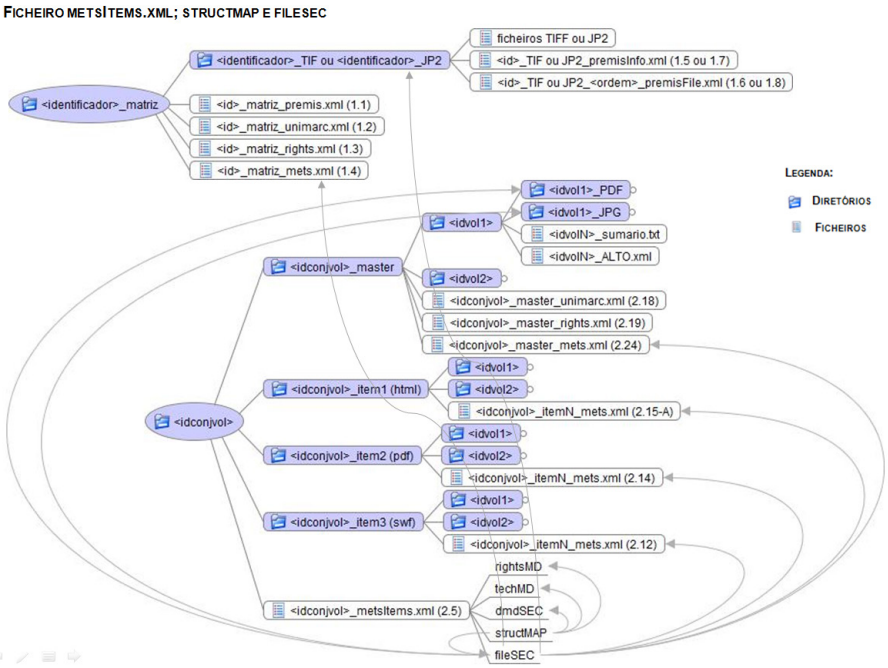

Figura 6 - Ficheiro metsItems: fileSec e structMap

O esquema PREMIS (http://www.loc.gov/standards/premis/v2/premis.xsd) foi selecionado para representar metadados técnicos e de preservação que suportam funções de manutenção de viabilidade, autenticação e identidade em contextos de preservação digital (PREMIS, 2012).

Na adaptação da versão 2.3 do esquema PREMIS aos objetos da BND, optou-se por aplicar apenas os elementos relativos às entidades Objeto e Direitos. Não se utilizam, assim, quaisquer elementos dos blocos 3 e 5, relativos às entidades Eventos e Agentes, uma vez que neste momento não estão ainda formalizadas para a BND ações de preservação digital que devam ficar documentadas, nem pessoas, organizações ou *software* associadas a essas ações. Os metadados técnicos e de preservação PREMIS estão armazenados nos objetos de informação na BND de acordo com a seguinte tipologia de ficheiros: ficheiros *premis.xml* (referenciam para preservação os pacotes de informação de disseminação), *premisInfo.xml* (descrevem as características técnicas não de cada imagem, mas de um mesmo grupo de conteúdos, por formato de ficheiro), *premisFile.xml* (contêm os metadados técnicos de cada ficheiro de conteúdo individualizadamente) e *premisItems.xml* (que identifica para preservação o ficheiro *metsItems.xml*, que empacota o objeto de informação na totalidade das suas partes componentes).

O esquema MIX, aprovado numa norma NISO (ANSI/NISO Z39.87. 2006) e disponível em http://www.loc.gov/standards/mix/mix20/mix20.xsd, aplica-se na BND apenas aos ficheiros de imagens matrizes, quer sejam TIFF ou JPEG 2000, com o objetivo de criar ficheiros de metadados que guardem, fora dos ficheiros de imagem, a informação técnica que usualmente está embebida nas etiquetas desses ficheiros, para além de representar outra informação técnica relevante que não conste neles. Os metadados MIX não constituem um ficheiro autónomo, estando antes integrados nos ficheiros *premisFile.xml*, embebidos no elemento PREMIS 1.5.7 "objectCharacteristicsExtension". Optou-se por esta implementação quer por se tratar de metadados técnicos fortemente ligados à preservação das matrizes do objeto, quer para evitar a proliferação de ficheiros já que, havendo um ficheiro de metadados *premisFile.xml* por cada ficheiro de imagem, seria excessivo estar a criar o

dobro de ficheiros ao invés de integrar esses metadados técnicos MIX nos *premisFile.xml* já existentes.

Para além do esquema MIX, foram adotados novos *standards* em substituição de esquemas locais, como é o caso do MarcXchange (ISO 25577. 2013), aplicado na representação dos metadados descritivos do recurso analógico digitalizado, cujo esquema (v.2.0, de julho 2013) está disponível em: http://loc.gov/standards/iso25577/marcxchange-2-0.xsd.

Aos metadados de direitos passou também a aplica-se o esquema *METSRights - The RightsDeclarationMD Extension Schema* (http://www.loc.gov/standards/rights/METSRights.xsd), em vez de um esquema local. O METSRights define um conjunto mínimo de metadados sobre direitos de propriedade intelectual associados ao objeto de informação. Conforme já referido, apesar de o esquema PREMIS conter elementos relativos a direitos, estes dizem respeito às atividades de preservação e não às permissões de acesso/distribuição.

Por último, não existindo no anterior modelo de dados representação *standard* do conteúdo resultante do reconhecimento ótico de caracteres (OCR), passou a aplicar-se o esquema ALTO - *Analysed Layout and Text Object* (http://www.loc.gov/standards/alto/alto-v2.0.xsd), por se tratar de um esquema mantido pela Biblioteca do Congresso como extensão do METS.

DZBTool - Ferramenta de produção e gestão dos objetos de informação

Para implementar o novo modelo de dados e o perfil de metadados que acima se apresentou, foi desenvolvida a ferramenta DZBTool que, para além da produção dos objetos de informação, permite gerir todo o respetivo processo de criação, desde o controlo inicial das imagens digitalizadas até ao registo no sistema PURL e ao arquivo dos objetos no armazenamento digital da BND, conforme fluxo de trabalho que se reproduz na Figura 7.

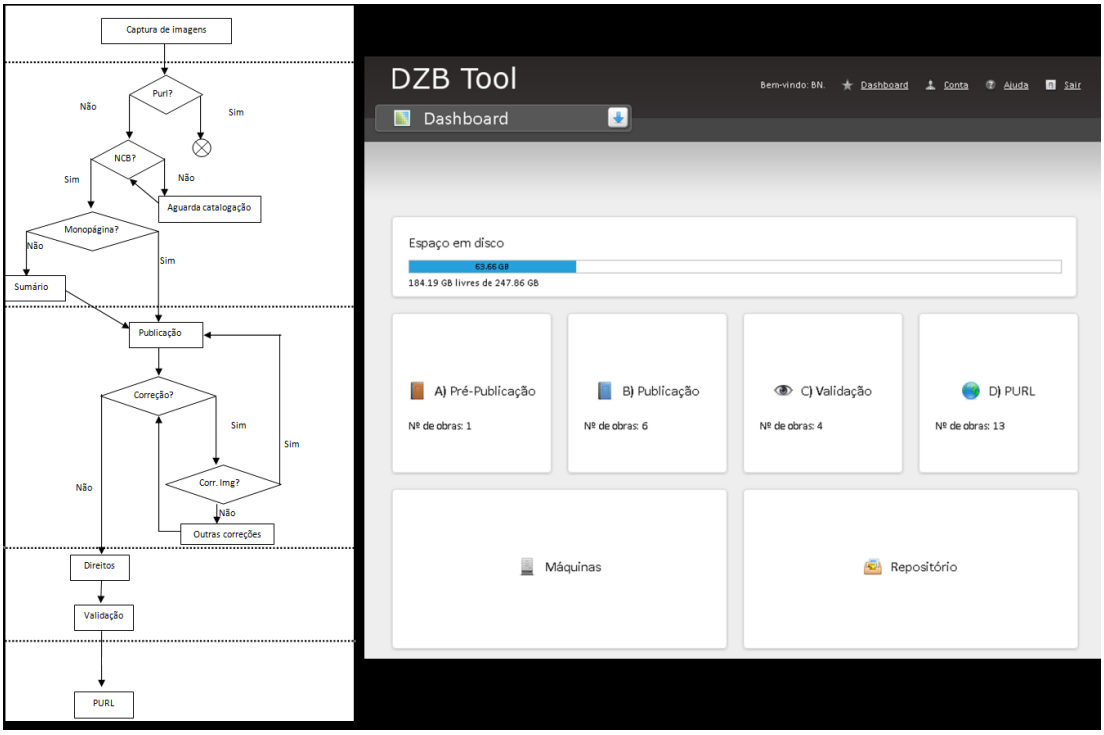

Figura 7 - Fluxo de trabalho publicação BND e Interface DZBTool

As fases e operações de trabalho não se alteraram, na sua essência, relativamente ao processo em uso desde 2007. Contudo, as diversas aplicações informáticas utilizadas estão agora integradas numa mesma ferramenta, a DZBTool, ou, em alternativa, com ela plenamente articuladas, conforme se desenvolverá em seguida.

Por outro lado, a DZBTool é, também, uma ferramenta de gestão do fluxo de trabalho, substituindo os anteriores instrumentos de controlo que tinham diferentes naturezas, não estavam integrados e dependiam em grande medida de inserção manual de dados. Passou, assim, a haver maiores garantias no controlo da sequência das operações, maior qualidade de dados e grandes ganhos de eficiência.

Relativamente à gestão dos processos, a DZBTool permite a monitorização das diferentes operações em tempo real, o que nos levou a abandonar instrumentos de controlo pouco eficientes, como por exemplo folhas Excel com dados introduzidos manualmente. A ferramenta possibilita também extrair relatórios de produção de forma mais automática e rigorosa e outros ganhos ao nível da gestão do processo, sobretudo, na recuperação do *backlog* da publicação, pois lista de forma automática as obras digitalizadas que aguardam processamento em determinado repositório. Por outro lado o controlo de duplicados é assegurado pela verificação automática da existência de registo da obra no sistema purl e pela eliminação automática dos repositórios de origem tanto dessas imagens já publicadas, como das imagens em curso de publicação no final do processo.

No que respeita à produção dos objetos, a integração das diferentes operações num único fluxo de trabalho automatizado impede a manipulação manual de ficheiros, o recurso a múltiplas aplicações para processamento de imagens, criação de sumários, processamento de metadados, criação de PDF, etc. Para um exemplo de integração de interfaces de diferentes operações ver Figura 8.

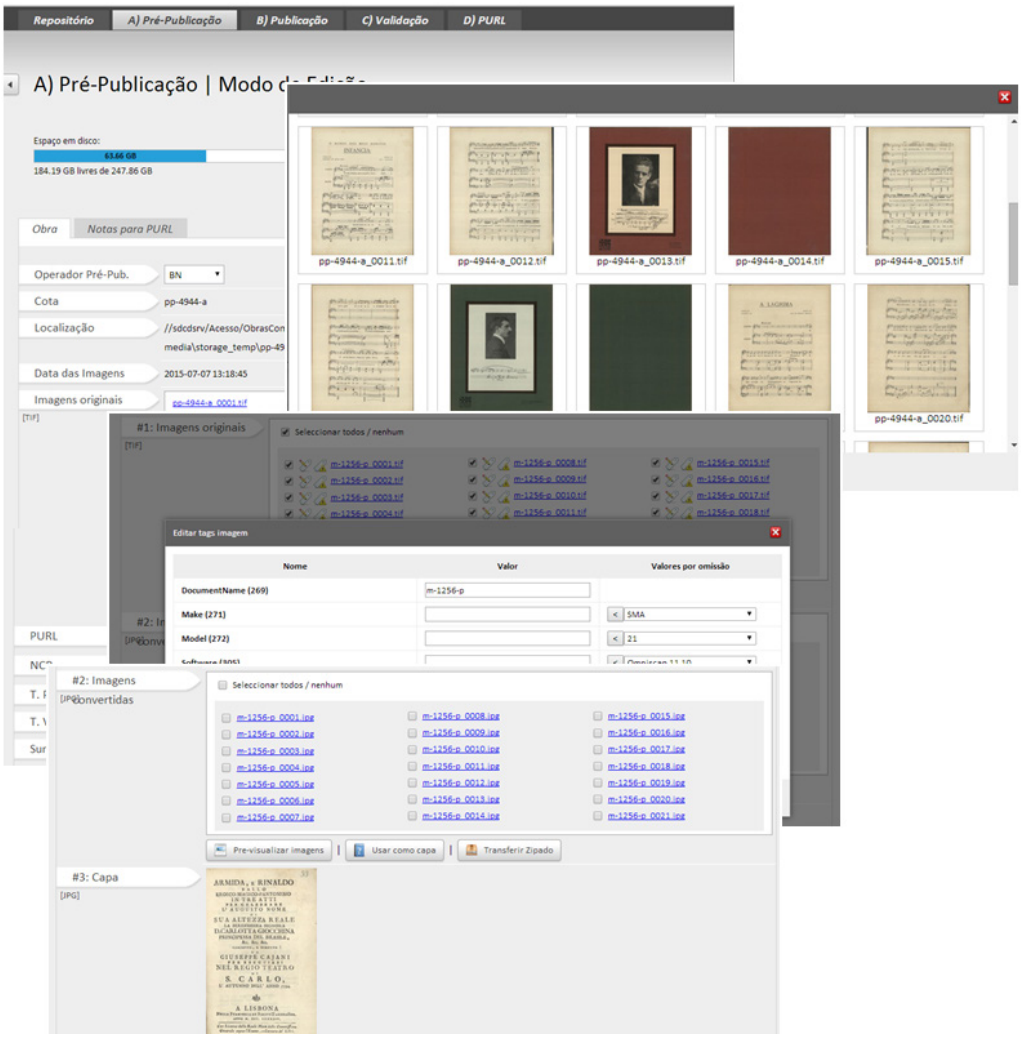

Figura 8 - DZBTool - Exemplo de controlo de imagens, edição de tags e criação de JPEG

Com a nova ferramenta passou a ser possível desenvolver todas estas operações sem sair da mesma aplicação, que regista de forma automática em que ponto do processo se encontra cada obra. Por exemplo, no mesmo interface é possível editar as etiquetas TIFF, verificar se as imagens estão

completas e na ordem correta, importar um sumário ou elaborá-lo no momento, criar os JPEG e enviar a obra para o LURA server, que executa o PDF e o OCR, continuando depois o processo de publicação no DZBTool. Isto sem embargo de se poder importar ficheiros JPEG, PDF ou sumários criados fora do sistema, sempre que tal situação seja mais vantajosa.

Há, ainda na produção, novas funcionalidades como a geração automática de *bookmarks* nos PDF, a partir dos sumários; a determinação dos direitos de autor na fase de validação da obra, com visualização das imagens, o que torna o processo muito mais célere; a possibilidade de indicar a data em que determinada obra cairá no domínio público (valor do elemento PREMIS «endDate») o que possibilitará a futura automatização da mudança do estatuto legal das obras em termos de direito de autor e a integração de correções e republicações de obras neste fluxo de trabalho, entre outras.

Conclusões

O novo modelo de informação da BND e o perfil de metadados definido para a sua implementação diferem do modelo ContentE por três ordens de razões principais: i) adoção de *standards* ou esquemas de metadados com ampla base de implementação, em detrimento de esquemas locais de metadados (os esquemas definidos localmente para os metadados técnicos, Unimarc e direitos deixaram de ser utilizados no novo modelo de metadados); ii) implementação de novos *standards* para representar metadados que anteriormente não existiam, nomeadamente através dos esquemas MIX, PREMIS e ALTO; correspondência entre o vocabulário local definido para metadados de direitos com o vocabulário estabelecido no PREMIS; simplificação dos ficheiros de metadados, sobretudo no caso dos ficheiros METS, em que se eliminaram redundâncias como a repetição de metadados técnicos por item e por imagem ou a indicação da ordem física e lógica de todos os ficheiros; e, iii) a não utilização de um ficheiro específico para meta-metadados relativos ao processo de criação de metadados, pois essa informação consta já dos ficheiros de preservação, técnicos e de estrutura.

A adoção destes modelos e *standards* internacionais, com as características acima descritas e com larga implementação nas comunidades de bibliotecas e arquivos digitais, permite assegurar a interoperabilidade dos dados da BND, disponibilizando-os em múltiplas plataformas internacionais e europeias (por ex., Europeana, The European Library, Biblioteca Digital do Património Iberoamericano, Biblioteca Digital Mundial), que propiciam quer a descoberta, acesso e reutilização dos nossos conteúdos, quer o relacionamento com outros recursos disponíveis nessas plataformas. O alinhamento com estas normas internacionais assegura, por outro lado, a gestão eficaz dos recursos digitais em todo o seu ciclo de vida, tendo em conta, desde logo, o impacto de futuras ações de preservação digital.

Com efeito, o novo modelo de informação da BND permite manter o nível de interoperabilidade dos dados, a integração em serviços coletivos e a partilha de dados em múltiplas plataformas, propiciando paralelamente uma estrutura de dados mais alinhada com os *standards* e modelos de referência internacionais, o que permite implementar estratégias e medidas concretas de preservação digital a longo prazo. O desenvolvimento de uma ferramenta simultaneamente integradora e modular, a par de uma estrutura de dados menos redundante, torna mais célere, ágil e eficiente todo o processo de produção, alteração e gestão do ciclo de vida de um objeto digital.

Referências bibliográficas

ALTO - *Analyzed Layout and Text Object*. Version 3.0, 2014. Disponível em: http://www.loc.gov/standards/alto/about.php.

ANSI/NISO Z39.87. 2006 - *Data Dictionary: Technical Metadata for Digital Still Images*.

BUCKLEY, Robert - *JPEG 2000: a practical digital preservation standard?* S.l.: Digital Preservation Coalition, 2008. Disponível em: www.dpconline.org/docs/reports/dpctw08-01.pdf .

CONSULTATIVE Committee for Space Data Systems – *Reference model for an Open Archival Information System (OAIS): magenta book.* Washington: CCSDS, 2012. Disponível em: http://public.ccsds.org/publications/archive/650x0m2.pdf.

CONSULTATIVE Committee for Space Data Systems – *Reference architecture for space information management: green book.* Washington: CCSDS, 2013.
Disponível em: http://public.ccsds.org/publications/archive/312x0g1.pdf.

DIGITAL Library Federation – *METS - Metadata Encoding and Transmission Standard: primer and reference manual.* Version 1.6. rev. Library of Congress**,** 2010.
Disponível em: http://www.loc.gov/standards/mets/METSPrimerRevised.pdf.

ENDERS, Markus - "A METS based information package for long term accessibility of Web archives". In *iPres 2010: Proceedings of the 7th International Conference on Preservation of Digital Objects*. Vienna: Osterreichische Computer Gesellschaft, 2010, pp. 31-39. Disponível em: http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/enders-70.pdf

FAUDET, Louise; PEYRARD, Sébastien – "A data-first preservation strategy: data management in SPAR." In *iPres 2010: Proceedings of the 7th International Conference on Preservation of Digital Objects*. Vienna: Oserreichische Computer Gesellschaft, 2010.
Disponível em: http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/fauduet-13.pdf

FOULONNEAU, Muriel; RILEY, Jenn - *Metadata for digital resources: implementation, systems design and interoperability*. Oxford, Chandos, 2008. 203 p. ISBN 978-1-84334-301-1.

HAHN, Matthias - *Recommendations for preservation-aware digital object model***.** Viena: SCAPE Project, 2014. Disponível em:
http://www.scape-project.eu/wp-content/uploads/2014/02/SCAPE_D8.1_FIZ_V1.0.pdf

ISO/IEC 11179-1 - 2004 - *Information tecnhology - Metadata registries*.

ISO 25577. 2013 - *Information and documentation - MarcXchange*.

ISO/TR 23081. 2009 - *Information and documentation - Managing metadata records*.

LAVOIE, Brian - *The Open Archival Information System Reference Model: introductory guide*. S.l.: OCLC, 2004. Disponível em: http://www.dpconline.org/docs/lavoie_OAIS.pdf.

LAVOIE, Brian; GARTNER, Richard - *Preservation metadata: technology watch report.* 2nd ed. York: Digital Preservation Coalition, 2013. Disponível em: http://dx.doi.org/10.7207/twr13-03

PATRÍCIO, Helena Simões - *Modelo de informação da Biblioteca Nacional Digital: modelo de dados; perfil de metadados*. Lisboa: BNP, 2015.

PEDROSA, Gilberto (2010) *ContentE: editor de conteúdos estruturados*. Versão 2.0. Disponível em: http://purl.pt/index/geral/PT/infoProfContentE.html#presentation.

PREMIS Editorial Committee – *PREMIS data dictionary for preservation metadata.* Version 2.2. Washington: PREMIS Editorial Commitee, 2012.
Disponível em : http://www.loc.gov/standards/premis/v2/premis-2-2.pdf.