View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Publicações BAD (Bibliotecários, Arquivistas e Documentalistas - E-Journals)

Depósito e Preservação na Biblioteca Nacional Digital

José Borbinha

Biblioteca Nacional – Direcção de Serviços de Inovação e Desenvolvimento
1749-081 Lisboa
Tel.: 217982083
E-mail: jose.borbinha@bn.pt

RESUMO
Este artigo aborda os problemas do depósito e preservação de informação digital na Biblioteca Nacional, tal como estão sendo encarados na perspectiva da iniciativa Biblioteca Nacional Digital. 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 em geral, os quais se têm mostrados suficientes no geral. Os novos desafios aqui abordados localizam-se a montante desses problemas, nas áreas das estratégias de depósito tendo em consideração os géneros e os actores envolvidos, e a jusante nas problemáticas do armazenamento e preservação desses conteúdos. Tudo isto implica a abordagem de problemas especiais de natureza técnica que são abordados neste artigo de uma forma acessível. O artigo apresenta ainda uma descrição breve dos principais casos de uso identificados para a Biblioteca Nacional Digital, assim como uma descrição da arquitectura tecnológica definida para os suportar.

PALAVRAS-CHAVE: Bibliotecas Digitais, Digitalização, Edição Digital, Depósito Digital, Preservação Digital.

INTRODUÇÃO

O problema do depósito de obras digitais em bibliotecas patrimoniais tem sido já motivo de várias abordagens, com alguns casos consequentes mas ainda sem um modelo universal generalizado.

De uma forma geral pode-se dizer que o problema tem três espaços:

- Conteúdos concebidos apenas para a Internet, os quais são regra geral criados no âmbito de processos que recorrem na sua globalidade a meios de produção digitais. Estão neste caso a grande maioria dos "sites" na Internet, alguns deles assentes em sistemas de publicação digital bastante complexos e sofisticados, como são os casos de alguns jornais diários de referência.

- Conteúdos concebidos para uma distribuição tanto na Internet como pelos meios tradicionais. Estes conteúdos são igualmente já produzidos regra geral sempre em ambientes digitais, embora de menor complexidade que o caso anterior. Exemplos a apontar são os casos de alguns documentos governamentais, relatórios, teses e dissertações (produzidos regra geral em ambiente chamados de "desktop"), publicações periódicas (especialmente revistas científicas), etc., distribuídos na Internet mas também impressos em papel.

- Conteúdos concebidos apenas para distribuição tradicional, como a impressão, mas em que pelo menos a parte final do processo é desenvolvida recorrendo exclusivamente a processos digitais (composição, revisão e impressão).

Se apontarmos a vulgaridade hoje em dia dos computadores pessoais (quase todos eles naturalmente equipados logo à partida com processadores de texto) e ainda o facto de a quase totalidade das tipografias já funcionarem igualmente em ambientes digitais, somos levados à interessante conclusão de que na realidade toda a informação e conteúdos que circulam hoje em dia na nossa sociedade existirem algures em formato digital. Levando esta constatação às suas últimas consequências, devemos então reconhecer que os problemas do depósito e preservação digital não devem ser encarados já como um excepção, mas talvez como a regra.

Este artigo aborda os problemas do depósito e preservação de informação digital 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. Estes casos compreendem 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 [17], os quais se têm mostrados suficientes no geral. Os novos desafios aqui abordados localizam-se a montante desses problemas, nas áreas das estratégias de depósito tendo em consideração os géneros e os actores envolvidos, e a jusante nas problemáticas do armazenamento e preservação desses conteúdos. Tudo isto levanta problemas especiais de natureza técnica que iremos abordar adiante, de uma forma que se tentou ser o mais acessível possível.

O artigo prossegue de seguida com uma descrição breve

dos principais casos de uso identificados para a BND, seguindo-se uma descrição da arquitectura tecnológica definida para os suportar. Esta arquitectura está neste momento em fase de desenvolvimento na BN, com vários componentes já em fase de exploração (uma primeira versão completamente funcional deverá estar pronta em Junho de 2004, com todos os componentes estabilizados até final do ano).

De seguida discute-se o problema do depósito digital em geral, com referências a casos noutras bibliotecas nacionais ou países. A esta discussão segue-se uma análise do problema da perspectiva nacional e referências à tecnologia que foi identificada como relevante para a abordagem do mesmo no âmbito da BND.

O artigo prossegue com uma descrição dos cenários de depósito digital em desenvolvimento para a BND, seguindo-se uma referência ao modelo de metadados estruturais seleccionado, e à tecnologia e processos de armazenamento desenvolvida para permitir a oferta de soluções de preservação.

BIBLIOTECA NACIONAL DIGITAL

Os principais casos, ou processos, considerados na iniciativa da BND são ilustrados na Figura 1, e brevemente descritos adiante.

Publicação

Considera-se a publicação como sendo o conjunto de todos os casos que resultam na criação de uma manifestação de uma expressão de uma obra para acesso por terceiros, independentemente dos seus formatos e modelos económicos (seja para efeitos privados, livre acesso, venda, licenciamento, etc.). Regra geral tal consiste num de dois cenários genéricos:

- Edição: A edição digital resulta na produção de obras originais em formato digital, incorporando ou não conteúdos digitalizados.

- Digitalização: A digitalização consiste regra geral na transcrição de obras impressas para formato digital (transcrição para imagem ou texto).

A própria BND comporta duas linhas de acção de criação de conteúdos através da edição digital e digitalização, perspectiva em que a BN actuará como qualquer outro produtor. A BND tem desenvolvido assim projectos de edição de títulos digitais originais, essencialmente obras de referência relacionadas com as colecções tradicionais da BN (e de preferência incluindo ou pelo menos referindo obras digitalizadas). Os objectivos das acções de digitalização são, para além de dar apoio aos projectos de edição, facilitar o acesso a obras de referência.

Aquisição

O caso da aquisição compreende as acções legais, comerciais e técnicas necessárias para que uma entidade possa entrar na posse física de uma publicação. Na BND estão sendo desenvolvidos processos e sistemas que permitem a constituição de espólios por aquisição de obras digitais ou digitalizadas, considerando os seguintes casos concretos:

- Recolha: A recolha é um processo de aquisição despoletado pela BN, aplicando-se a obras ou recursos publicados em geral na Internet. Estão sendo desenvolvidos sistemas de recolha automatizada de recursos da Internet.

- Depósito: O depósito de obras na BN poderá efectuar-se por obrigação legal, contratual ou de modo voluntário. Estão sendo desenvolvidos serviços de

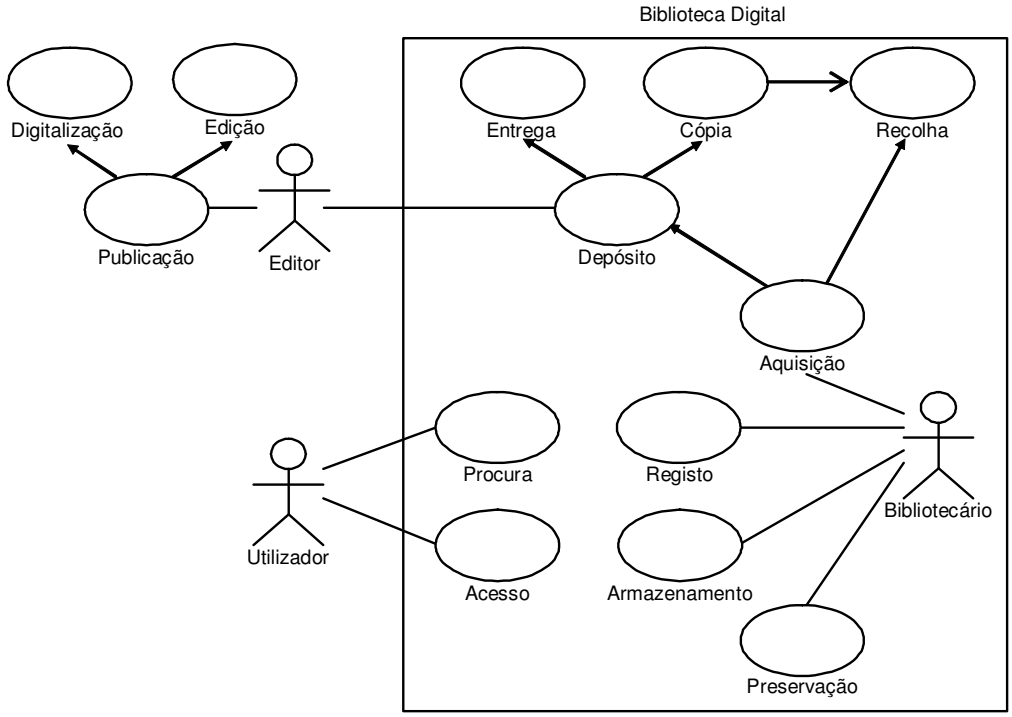

Figura 1: Casos na Biblioteca Nacional Digital

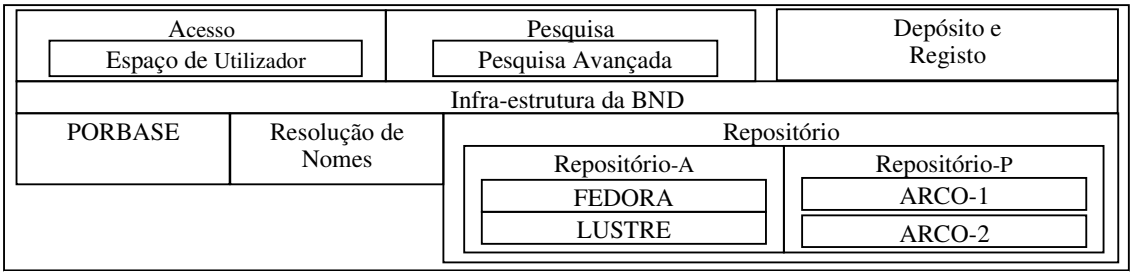

Figura 2: Arquitectura da BND.

depósito de obras pela Internet, a executar segundo dois cenários principais:

- Entrega: A entrega corresponde a uma forma de depósito efectuada em tempo real, durante a qual a obra é explícita e efectivamente entregue à BN (a entidade de depósito).

- Cópia: A cópia corresponde ao caso em que se regista a intenção de depósito, mas o mesmo só será realmente efectuado num momento posterior, através de um mecanismo de recolha despoletado pela entidade depositária.

Registo
O registo compreende o conjunto de tarefas relacionadas com a recolha, organização e gravação de informação sobre uma publicação, necessária para a sua descrição, gestão, procura, acesso e preservação (tal como informação bibliográfica, de licenciamento, etc.).

Será garantido o registo bibliográfico na PORBASE de todas as obras da BND, com a possibilidade do registo de informação complementar noutros sistemas especializados.

Procura
A procura refere-se aos processos de descoberta das publicações que respondam a um determinado conjunto de critérios (pesquisa bibliográfica, navegação em índices, etc.). Podem-se incluir neste caso cenários de disseminação selectiva de informação.

Será oferecida a possibilidade de pesquisa e descoberta de todas as obras da BND através da PORBASE, assim como através de sistemas derivados tais como índices especializados ou cópias parciais da PORBASE (para facilitar pesquisas especializadas). Serão desenvolvidas ainda soluções para serviços complementares de pesquisa exteriores, oferecendo interoperabilidade por exemplo para portais e motores de pesquisa.

Acesso
O acesso é o processo que garante a um determinado utilizador usufruir de uma publicação, na forma de um item ou conjunto de itens que definem o conteúdo intelectual desse recurso.

Será garantido o acesso às obras da BND, nos regimes de acesso livre (preferencial) ou acesso controlado para recursos com características especiais. Serão oferecidos ainda serviços de acesso a conteúdos em suportes físicos, tais como acesso em CD-ROM ou DVD para obras digitalizadas (que poderão ser enviados para casa dos leitores, contendo cópias de obras digitalizadas com imagens digitalizadas em alta resolução, com melhor qualidade portanto das cópias acessíveis na Internet).

Armazenamento
O armazenamento compreende todos os processos e técnicas destinadas a colocar uma publicação num local onde a mesma possa ser acedida (segundo as suas regras específicas ou outras gerais).

Será construído um sistema de armazenamento de grande capacidade, escalável, tolerante a falhas, e de baixo custo relativo. Este sistema incluirá um espaço de acesso, na rede externa da BN, e outros de segurança, na rede interna da BN ou ainda noutros locais.

Preservação
A preservação compreende todos os processos destinados a garantir a boa forma física, lógica e intelectual da publicação, tendo como objectivo último garantir o seu acesso em qualquer momento futuro.

Nesta fase da BND será dada uma atenção especial à preservação física de longo prazo dos conteúdos.

Arquitectura da BND
Tendo em vista fornecer o necessário suporte técnico aos casos atrás descritos, foi concebida para a actual fase da BND a arquitectura representada na Figura 2.

Será de seguida fornecida uma descrição breve para cada um dos seus componentes, sendo os dois componentes realçados na Figura 3 abordados com mais detalhe funcional no resto do artigo.

O sistema de Depósito e Registo suporta todos os processos respectivo à própria designação, sendo abordado com detalhe mais adiante.

O sistema de Pesquisa oferece acesso à PORBASE ou ainda a índices ou bases de dados bibliográficas especializadas. Uma versão avançada deste sistema irá oferecer funções inovadoras de pesquisa avançada, combinando os metadados descritivos (registos bibliográficos) com a análise completa dos textos dos conteúdos e ainda da sua estrutura (ver descrição sobre os metadados estruturais adiante).

O sistema de Acesso permitirá o registo de utilizadores, e a oferta de serviços avançados no Espaço de Utilizador, tais como a criação de colecções virtuais próprias (que ficarão registadas entre acessos, e das quais poderão ainda ser

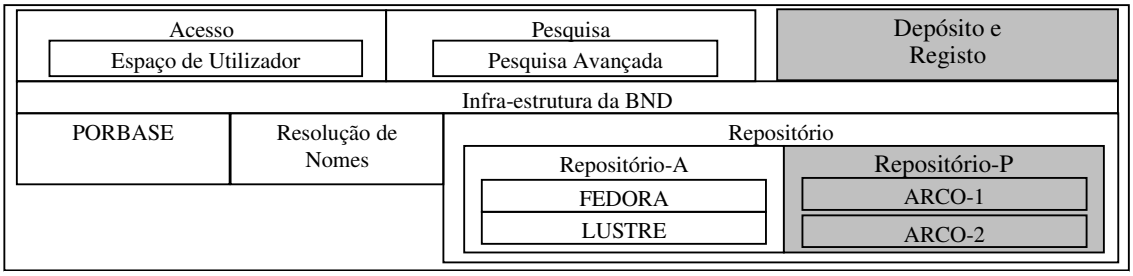

Figura 3: Realce aos componentes detalhados funcionalmente neste artigo.

solicitadas cópias em CD-ROM ou DVD, que serão fornecidos ao leitor com os conteúdos para os quais haja autorização para tal).

Os três sistemas de serviços operam sobre uma Infra-estrutura, a qual permite a comunicação com um nível mais baixo de sistemas onde se inclui a própria PORBASE (especialmente para registo e acesso a informação sobre exemplares) e ainda um sistema de Resolução de Nomes. De momento o sistema de Resolução de Nomes é composto apenas por um serviço de resolução de apontadores persistentes [21], mas a breve prazo consistirá numa solução mais completa, com um serviço de resolução genérica de apontadores OpenURL [16].

Finalmente temos o sistema de Repositório, composto por componentes de código aberto (servidores Linux e aplicações desenvolvidas por indicação da BN ou em projectos externos). Este é composto ainda por dois subsistemas, um Repositório-A para acesso, e outro Repositório-P para preservação. O Repositório-A é composto por uma plataforma FEDORA [6] de suporte a serviços, e ainda por um sistema de ficheiros escalável e de grande capacidade LUSTRE [11]. O modelo funcional do sistema Repositório-P está relacionado com os objectivos de preservação da BND, sendo por isso abordado com mais algum detalhe mais adiante.

DEPÓSITO DIGITAL: ALGUMAS REFERÊNCIAS

O depósito de recursos digitais em suporte físico (CD-ROM) é já prática corrente em muitos países, especialmente quando se trata de artefactos com conteúdos de alguma forma equivalentes a monografias ou publicações periódicas.

A grande maioria das acções mais recentes e relevantes levadas a cabo pelas bibliotecas nacionais a nível internacional tem sido no domínio da Internet. Existem duas abordagens principais na recolha de conteúdos da Internet: a recolha indiferenciada de um ou mais domínios (como por exemplo o espaço *.pt), ou a recolha selectiva de "sites" ou grupos.

O primeiro caso tem o interesse de, ao se recolher automaticamente todo um espaço, se minimizar o esforço humano envolvido no processo. Como aspecto negativo pode haver o facto de não serem criados metadados descritivos dos recursos (pelo menos numa primeira fase), assim como o facto de possivelmente se ter de destinar espaço de armazenamento para conteúdos irrelevantes.

A recolha com base numa selecção individual ou em grupos dos recursos tem como vantagens permitir a selectividade e a criação imediata de pelo menos alguns metadados descritivos dos recursos depositados, o que potenciará vir-se a oferecer melhores serviço de pesquisa. Como inconveniente há o custo do esforço humano, que pode ser considerável, e ainda o risco de se poder vir a deixar de fora recursos importantes.

Exemplos de alguns projectos ou iniciativas nestas áreas são:

- Depósito Selectivo de Recursos Estáticos: A Dinamarca [18] e o Canadá [2] são os principais exemplos desta abordagem. Recursos que podem ser equiparados a publicações impressas e que não mudam nem contêm elementos dinâmicos ou interactivos são arquivados com uma base selectiva (decidida pela biblioteca de depósito).

- Depósito Selectivo de Recursos da Internet Estáticos ou Dinâmicos: O projecto PANDORA [19], na Austrália, é um exemplo do arquivo de publicações e sítios dinâmicos ou estáticos, mais uma vez com base na intervenção e trabalho intelectual da parte da biblioteca de depósito.

- Depósito de Domínios da Internet: Algumas bibliotecas ou organizações equivalentes tentam fazer recolha automática de todo o domínio da Internet dos seus países respectivos usando robots d e um mínimo de intervenção humana (envolvendo geralmente ainda recursos relacionados existentes noutros domínios genéricos, como ".com", ".net", ".org", etc.). Existe um projecto nesse sentido no Reino Unido [20], em conjunto com a iniciativa Internet Archive [9], assim como um outro da associação de bibliotecas nórdicas Nordic Web Archive [14].

- Depósito Voluntário: A Biblioteca Real da Holanda [5] desenvolveu uma infra-estrutura técnica e relação organizacional com alguns editores, para arquivar, preservar e fornecer acesso limitado a tudo o que foi produzido pelos mesmos em formato digital.

Para a BND decidiu-se dar preferência nesta fase a estratégias de recolha de conteúdos despoletadas por processos de selecção ou de propostas para depósito, podendo ainda incidir sobre sítios com múltiplas características técnicas, legais ou outras. A recolha de sítios tem de ser assim resolvida de forma diferente consoante o cenário em que se enquadre.

Relativamente às técnicas básicas de recolha do depósito elas irão seguir dois modelos gerais, um baseado em

robots de recolha e outro em processos de sincronização entre o sítio da publicação e o depósito. Além disso há que levar ainda em linha de conta factores de ordem temporal (quando é um recurso publicado ou modificado).

Uma vez que os técnicos que vão ter de configurar cada caso podem não ter conhecimento da tecnologia usada na criação ou publicação do recurso, todos os cenários precisam de uma fase posterior de verificação, para resolver problemas ou erros, ou para alterar dados de carácter tecnológico.

Independentemente da forma como for feita a recolha dos recursos, deve-se ter ainda em consideração que os meios a utilizar serão sempre limitados. Os meios fundamentais que vão estar em causa serão o espaço para armazenamento e a largura de banda usada. Estes meios vão ser ainda partilhados com diversas aplicações e serviços na BN, pelo que esse requisito terá de ser também considerado na solução.

Formas de Acção

A BND concentra-se no problema do depósito selectivo. Todos os recursos a considerar para depósito serão assim prévia ou posteriormente verificados. O processo de depósito poderá no entanto ser desencadeado directamente por actores dedicados a essa tarefa (funcionários ou colaboradores da BN) ou após sugestões de terceiros. Tal permite definir os dois grandes cenários:

- Depósito Unilateral: Este caso dá-se quando se pretende depositar uma publicação sem que haja um entendimento formal com o autor/responsável dessa publicação para tal. Este caso é específico de sítios em que não é necessário interagir com esse autor ou responsável para se poder fazer a cópia do sítio. Como consequência, este depósito não implica nenhum esforço adicional da parte dos donos do sítio, para além da utilização dos seus recursos por parte da BN para a cópia. Naturalmente, no caso de sítios grandes, pode ser problemático se a Biblioteca Nacional tentar armazenar todo o conteúdo do sítio. Neste cenário o processo de depósito é efectuado exclusivamente pela BN, sem o envolvimento directo dos agentes responsáveis pela criação ou publicação do recurso.

- Depósito Bilateral: Quando a BN conseguir entrar em acordo com os responsáveis por uma publicação, para depósito do conteúdo de forma sincronizada e sistemática dessa publicação, o trabalho de tratamento do conteúdo dessa publicação fica facilitado pois pode ser feito do lado de quem publica, possivelmente com muito menos esforço do que a para a BN. A transmissão também pode ser facilitada por intermédio de uma solução tecnológica que seja proveitosa para os dois lados. Quem publica também pode ter interesse em ter as publicações armazenadas de forma estruturada e facilmente acessível sem ter de despender nenhum esforço adicional para isso. Neste caso pressupõe-se a existência de um acordo entre as duas partes (a BN e o editor), o que pode permitir optimizar alguns passos do processo ou ultrapassar obstáculos (como autenticação).

Géneros dos Recursos

De uma forma geral assume-se que todos os recursos a depositar poderão vir a ser alterados no tempo, mesmo aqueles que se julguem estáticos (o que acontece frequentemente para correcções pontuais de erros). As implicações disso poderão no entanto variar de caso para caso, com adiante se discute. Do ponto de vista dos géneros irão ser consideradas especialmente as seguintes quatro classes de recursos:

- Recursos Estáticos: Consideram-se neste caso os recursos que à partida não são modificados no tempo, a não ser para pequenos ajustes esporádicos. Neste caso o recurso deverá ser depositado em versões, considerando-se cada versão como sendo diferente da anterior. Exemplos são sítios WEB de referência, com alterações esporádicas ou mesmo nunca existentes.

- Recursos Dinâmicos: Os recursos dinâmicos são aqueles que produzem um resultado diferente em cada acesso. Os sítios dinâmicos podem ter páginas geradas de forma sempre diferente para cada acesso, com conteúdo modificado ou não, sendo por isso complexo descobrir quando é que realmente mudaram. Além disso é preciso usar tecnologia que consiga captar a maior parte da informação, o que pode ser complicado nestes casos. Casos como por exemplo os resultados de motores de pesquisa não são passíveis de armazenamento porque podem não ter limite de possibilidades.

- Recursos Periódicos: Alguns recursos podem ser de actualização periódica bem definida (publicações periódicas) com modificações de conteúdos e estrutura, que podem ir de segundos a horas ou meses, consoante o género. Essa frequência de actualização é no entanto conhecida, sendo levada em conta para o depósito. Neste caso o depósito consiste apenas numa versão do recurso, a qual é assumida à partida como sendo cumulativa em cada alteração. Deverá no entanto ser previsto para estes casos um mecanismo que permita em qualquer momento passar a registar uma nova versão do recurso se num dado instante este sofrer uma alteração estrutural significativa. Exemplos destes recursos são os jornais e revistas.

- Recursos Não Periódicos: Estão neste caso os recursos de actualização frequente, mas de periodicidade não definida. Estes recursos podem ter alterações de conteúdo e estrutura muito pequenas, com ligeiras alterações ou incrementos de cada vez, ou podem ter alterações bastante drásticas. Tal como no caso anterior, o depósito consistirá apenas numa versão do recurso, a qual é assumida à partida como sendo cumulativa em cada alteração, devendo no entanto ser previsto um mecanismo que permita em qualquer momento passar a registar uma nova versão do recurso.

Tecnologia para Depósito Digital na Internet

Independente à partida dos cenários descritos, existe ainda tecnologia à qual podemos recorrer, adaptando-a às necessidades ou simplesmente reutilizando-a. Serão usadas essencialmente quatro ferramentas para lidar com os vários cenários em que se enquadram as publicações digitais (note-se que as quatro ferramentas são de código

Cenário	Casos	Tecnologia		
		HTTRACK	**LOCKSS**	**RSYNC ou UNISON**
C1	DLib	(X)	(X)	Bilateral, Periódico
C2	Colecção Gutenberg	(X)	(X)	Bilateral, Não Periódico
C3	Disputatio	(X)	Bilateral, Periódico	
C4	Partidos Políticos	Unilateral, Não Periódico		
C5	Autores Portugueses	Unilateral, Estático		
C6	Blogs	Unilateral, Não Periódico		
C7	Publicações Periódicas	Unilateral, Periódico	(X)	
C8	Estáticos Variados	Unilateral, Estático		
C9	Não Periódicos Variados	Unilateral, Não Periódico		

Tabela 1: Cenários para depósito na Internet

aberto):

- HTTrack [8]: Este é um programa que tem como objectivo recolher e guardar páginas da Internet da forma mais fielmente possível parecidas com as originais que se encontrem no servidor. Para tal, faz a análise e interpretação não só de HTML mas também de diversos formatos e linguagens usados em páginas (como Javascript, Perl, VBScript, Java, Flash, etc.). Com este programa é possível configurar vários parâmetros para controlar a cópia (limitação dos domínios da Internet a que se pretende aceder, profundidade dos elos que serão guardados e tamanho máximo dos ficheiros, possibilidade de parar a cópia e de a retomar mais tarde, escolha do tipo de ficheiros que se pretende guardar, e ainda a possibilidade de criar filtros com expressões regulares). É um sistema útil tanto para projectos com o da BND como para utilizadores individuais.

- LOCKSS [10]: O projecto LOCKSS tem como objectivo a criação de uma rede de bibliotecas que partilhem entre si o depósito de publicações científicas acessíveis na Internet. Parte-se do pressuposto que existe um acordo entre a editora da publicação a copiar e a biblioteca que o pretende fazer, o que facilita os processos legais de partilha entre os servidores que vão guardar as cópias das diversas publicações. Existem neste projecto algumas particularidades não muito comuns, como a segurança e a partilha de dados e comunicação. O projecto está ainda em desenvolvimento, sendo a BN uma das bibliotecas pioneiras nos testes e acompanhamento.

- RSYNC [22]: Esta tecnologia tem como objectivo a transferência incremental rápida de ficheiros, ou seja, serve para acelerar a transmissão de ficheiros entre máquinas remotas. No caso de a cópia se ter efectuado pelo menos já uma vez, esta tecnologia permite efectuar actualizações transmitindo apenas as partes dos ficheiros que tenham sido modificadas, o que permite poupar tempo na transmissão.

- UNISON [23]: Esta tecnologia é baseada no mesmo algoritmo do RSYNC para comparação e transferência de ficheiros remotos. A sua utilidade principal será para a sincronização de conteúdos em máquinas com sistemas operativos Windows, o que não é suportado pelo RSYNC (o qual é no entanto uma ferramenta mais estável e madura para os outros casos, especialmente para sistemas Unix e Linux). Esta ferramenta está ainda orientada para sincronização bi-direccional

CENÁRIOS DE DEPÓSITO DIGITAL NA BND

O problema irá ser abordado por três grupos de cenários, um para cada um dos espaços referidos no início deste artigo.

Recursos da Internet

Para o espaço da Internet não são considerados neste momento cenários de recursos dinâmicos. No entanto irá ser deixada a possibilidade em aberto, na expectativa de que alguns resultados neste momento não previstos possam vir a possibilitar abordar pelo menos um exemplo.

Cada cenário da Internet é caracterizado pela instanciação de três factores: forma de acção; géneros dos recursos; tecnologia utilizada. Da conjugação destes três factores resulta um cenário de trabalho. Os cenários a ser resolvidos são (Tabela 1):

- Cenário 1 – Depósito Bilateral Periódico com Rsync: Este caso compreende o depósito de recursos do género periódico em que os responsáveis tenham uma participação activa no processo. O principal caso a testar será a revista DLIB [24].

- Cenário 2 – Depósito Bilateral Não Periódico com Rsync: Este caso compreende o depósito de recursos do género não periódico em que os responsáveis tenham uma participação activa no processo. O principal caso a testar será a colecção Gutenberg [25].

- Cenário 3 – Depósito Bilateral Periódico com LOCKSS: Este caso compreende o depósito de recursos do género periódico que autorizem a recolha e partilha entre várias entidades depositárias. Um exemplo a testar será a revista Disputatio [26].

- Cenário 4 – Depósito Unilateral de Grupos Uniformes Não Periódicos com HTTrack: Este caso compreende o depósito de recursos do género não periódico e com actualizações pouco frequentes, ou com picos de frequência de actualização. O principal caso a testar será a lista dos partidos políticos em Portugal [27].

- Cenário 5 – Depósito Unilateral de Grupos Estáticos com HTTrack:. Este caso compreende o depósito de recursos do género estático O principal caso a testar será uma base de dados de sites sobre Autores Portugueses na Internet (em desenvolvimento neste momento).

- Cenário 6 – Depósito Unilateral de Grupos Não Periódicos com HTTrack:. Este caso compreende o depósito de recursos do género não periódico e com

elevada frequência de actualização. O principal caso a testar será uma lista de "blogs" relacionados com Portugal [28].

- Cenário 7 – Depósito Unilateral de Grupos Periódicos com HTTrack:. Este caso compreende o depósito de recursos do género periódico. O principal caso a testar será uma lista de publicações periódicas portuguesas ou relativas a Portugal.

- Cenário 8 – Depósito Unilateral de Recursos Estático com HTTrack. Este caso compreende o depósito de recursos variados do género estático.

- Cenário 9 – Depósito Unilateral de Recursos Não Periódicos com HTTrack. Este caso compreende o depósito de recursos variados do género não periódicos.

Obras Concebidas para Impressão ou Distribuição na Internet

O caso das obras concebidas para impressão e ainda possível distribuição na Internet será abordado em dois cenários complementares:

- Iniciativa DiTeD – Teses e Dissertações [3]: Nesta iniciativa irá procurar-se a colaboração das bibliotecas universitárias para a criação de uma rede nacional de depósito de teses e dissertações digitais. O objectivo é garantir o depósito dessas obras localmente nas bibliotecas ou universidades, com uma cópia na BN. Para o efeito a BN desenvolveu uma solução informática adequada para a gestão de um repositório local, disponível em código aberto para quem a pretender utilizar como sistema de gestão local.

- Monografias Digitais: Para outras monografias será desenvolvido um serviço de depósito voluntário na Internet através do qual qualquer entidade, pública ou provada, poderá submeter as suas monografias, as quais serão catalogadas na PORBASE e armazenadas tal como as monografias impressas.

Obras concebidas para Impressão

Uma outra interessante linha de actividade é o depósito pelos editores de cópias digitais de obras impressas. O processo através do qual se propõe levar a cabo este objectivo é através do estabelecimento de parcerias com os editores, sensibilizando-os para o mútuo interesse nesse depósito. Efectivamente, através da adesão a esta proposta os editores poderão passar a ter acesso a um serviço que servirá também como segurança para os próprios, permitindo-lhes em qualquer altura recuperar cópia das suas obras aditadas.

É de esperar que este serviço verifique algumas reticências iniciais da parte de muitos editores, relacionadas com o facto de essas obras poderem desta forma vir a sair do seu controlo. Julga-se que essas dúvidas serão no entanto ultrapassadas quando a BN for capaz de demonstrar que consegue manter armazenadas em condições de segurança satisfatórias as obras que lhe forem entregues, respeitando assim os direitos devidos aos respectivos editores.

O depósito destas obras na BN, para além da sua preservação em si, poderá ainda vir a permitir uma exploração vantajosa das mesmas através da sua indexação global e utilização dos índices resultantes em serviços avançados de pesquisa, complementares à PORBASE. Os resultados dessas pesquisas poderão assim levar o utilizador a aceder imediatamente à obra, se esta estiver disponível para tal na BND, ou à apresentação da informação sobre os locais e formas de o conseguir em alternativa (como por exemplo o serviço encomendas pela Internet do próprio editor).

REPOSITÓRIO DE PRESERVAÇÃO

O serviço construído na BND para repositório de preservação dos conteúdos assenta na selecção de um modelo de metadados estruturais e numa solução informática em código aberto, facilmente escalável e de custo relativamente reduzido para grandes quantidades de informação.

Metadados Estruturais

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 digital do mesmo livro. Sem metadados estruturais, os ficheiros com imagens ou texto que compõem a obra digital 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 acesso a metadados técnicos apropriados para lhe permitir refrescar e migrar os dados, garantindo a durabilidade de recursos valiosos.

O projecto MOA2 – Making of America II [13] tentou abordar parcialmente estas questões providenciando um formato de codificação para metadados administrativos e estruturais para trabalhos textuais e baseados em imagens. Uma iniciativa da DLF [4] continuou esse trabalho, tendo desenvolvido o formato METS [12], um formato em XML para codificar metadados necessários tanto para a gestão de objectos de bibliotecas digitais num repositório como para a troca desses objectos entre repositórios.

Um documento METS consiste em sete secções principais:

- Cabeçalho: O cabeçalho contém metadados descrevendo o documento METS em si, incluindo informação como o criador, editor, etc.

- Metadados Descritivos: A secção de metadados descritivos pode apontar para metadados descritivos externos ao documento METS (e.g., um registo UNIMARC ou um registo EAD acessíveis num servidor na Internet), ou conter metadados descritivos

embebidos, ou ambos. Múltiplas instâncias de metadados descritivos, tanto internas como externas, podem ser incluídos na secção de metadados descritivos.

- Metadados Administrativos: A secção de metadados administrativos 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, e informação sobre a proveniência dos ficheiros que compõem o objecto digital (i.e., relações de ficheiros originais ou derivados, e informação de migração ou transformação). Tal como os metadados descritivos, os metadados administrativos podem ser tanto externos ao documento METS, ou codificados internamente.

- Secção de Ficheiros: A secção de ficheiros lista todos os ficheiros que contêm as versões electrónicas do objecto digital. Elementos do tipo ficheiro podem ser agrupados em elementos de grupos de ficheiros, para permitir a subdivisão de ficheiros por versão do objecto.

- Mapa Estrutural: O Mapa Estrutural é o coração do documento METS. Ele esboça uma estrutura hierárquica para o objecto da biblioteca digital, e liga os elementos dessa estrutura a ficheiros com conteúdos e metadados referentes a cada elemento.

- Ligações Estruturais: A secção de Ligações Estruturais do METS permite registar a existência de 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 tem um elemento de definição de interface que representa uma definição abstracta do conjunto de comportamentos para uma secção em particular, e ainda um elemento que identifica um módulo de código executável que implementa e executa esses comportamentos.

O esquema METS oferece um mecanismo flexível para codificar metadados descritivos, administrativos e estruturais para um objecto de uma biblioteca digital, e para exprimir as ligações complexas entre estas várias formas de metadados. Assim o 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 as principais funcionalidades do esquema, mas uma examinação mais detalhada do esquema e da sua documentação é necessária para compreender todo o alcance das suas capacidades.

Dependendo da sua utilização, um documento METS pode ser usado como Pacote de Informação de Submissão (*Submission Information Package* - SIP), um Pacote de Informação de Arquivo (*Archival Information Package* - AIP) ou um Pacote de Informação de Disseminação (*Dissemination Information Package* - DIP) no contexto do modelo de referência do OAIS – Open Archival Information System [15].

Todas as obras depositadas na BND serão estruturadas segundo o esquema METS, e enviadas para o repositório nessa forma.

Tecnologia ARCO

A peça chave do repositório de preservação da BND é a tecnologia de código aberto ARCO [7], desenvolvida pela ADDETTI.

Esta solução assenta em tecnologia Linux, permitindo construir volumes de armazenamento de grande capacidade com equipamentos heterogéneos. Cada volume de armazenamento pode ser composto por um número qualquer de servidores, ou nós, podendo cada nó ter qualquer número de discos e quantidade de espaço de armazenamento.

Com esta solução é possível configurar modelos de tolerância a falhas robustos, garantindo-se por exemplo que em cada volume cada obra está sempre armazenada em mais do que um disco, e que a sua recuperação é sempre possível em cenários de falha completa de um disco ou mesmo de um servidor.

Para aumentar as garantias, é ainda possível instalar e configurar remotamente (em local fora da BN) cópias dos volumes de preservação da BND. Para o efeito irá a BN desencadear um processo para a identificação de uma entidade parceira, com quem venha a ser possível negociar a instalação desses volumes de réplica.

NOTAS FINAIS

O trabalho relatado neste artigo tem sido desenvolvido na BN desde o início de 2003, embora com base num vasto leque de experiências levadas anteriormente a cabo. A fase actual do projecto está prevista terminar no final de 2004, e conta com o apoio do POSI – Programa Operacional da Sociedade da Informação. As acções de digitalização contarão ainda em 2004 com o apoio do POC – Programa Operacional da Cultura. As contrapartidas da BN são asseguradas pelos seus orçamentos próprios (Orçamento de Estado e PIDDAC). O investimento total previsto para todas estas acções deverá rondar no final dos dois anos um milhão de Euros.

REFERÊNCIAS

1. BND – Biblioteca Nacional Digital [Em linha]. [Consult. 29 Fev. 2004]. URL: <http://bnd.bn.pt>
2. Colecção Electrónica (Canadá) [Em linha]. [Consult. 29 Fev. 2004]. URL: http://collection.nlc-bnc.ca/e-coll-e/about-e.htm
3. DiTeD – Dissertações e Teses Digitais [Em linha]. [Consult. 29 Fev. 2004]. URL: http://dited.bn.pt
4. DLF – Digital Library Federation [Em linha]. [Consult. 29 Fev. 2004]. URL: http://www.diglib.org/dlfhomepage.htm
5. e-Depot [Em linha]. [Consult. 29 Fev. 2004]. URL: http://www.kb.nl/kb/resources/frameset_kb.html?/kb/menu/ken-arch-en.html
6. FEDORA – Flexible Extensible Digital Object and Repository Architecture [Em linha]. [Consult. 29 Fev. 2004]. URL: http://www.fedora.info/
7. Han-fei, et all – ARCO: Moving digital library storage to grid computing. ICEIS 2004, 6th International

Conference on Enterprise Information Systems, 14-17 April 2004, Porto, Portugal.
8. HTTrack [Em linha]. [Consult. 29 Fev. 2004]. URL: http://www.httrack.com
9. Internet Archive [Em linha]. [Consult. 29 Fev. 2004]. URL: http://www.archive.org
10. LOCKSS [Em linha]. [Consult. 29 Fev. 2004]. URL: http://lockss.stanford.edu/
11. LUSTRE – Scalable Clustered Object Storage [Em linha]. [Consult. 29 Fev. 2004]. URL: http://www.lustre.org/
12. METS – Metadata Encoding and Transmission Standard [Em linha]. [Consult. 29 Fev. 2004]. URL: http://www.loc.gov/standards/mets/
13. MOA2 - The Making of America II [Em linha]. [Consult. 29 Fev. 2004]. URL: http://sunsite.berkeley.edu/MOA2/
14. Nordic Web Archive [Em linha]. [Consult. 29 Fev. 2004]. URL: http://nwa.nb.no
15. OAIS – Reference Model for an Open Archival Information System. [Em linha]. [Consult. 29 Fev. 2004]. URL: http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html
16. OpenURL – The OpenURL Framework for Context-Sensitive Services. [Em linha]. [Consult. 29 Fev. 2004]. URL: http://www.niso.org/committees/committee_ax.html
17. PORBASE – Base Nacional de Dados Bibliográficos [Em linha]. [Consult. 29 Fev. 2004]. URL: http://www.porbase.org
18. Projecto Electra [Em linha]. [Consult. 29 Fev. 2004]. URL: http://www.kb.dk/elib/index-en.htm
19. Projecto Pandora [Em linha]. [Consult. 29 Fev. 2004]. URL: http://www.nla.gov.au/pandora/
20. Projecto WebArchive [Em linha]. [Consult. 29 Fev. 2004]. URL: http://www.pro.gov.uk/webarchive/
21. PURL.PT – Serviço de resolução de apontadores persistentes [Em linha]. [Consult. 29 Fev. 2004]. URL: http://purl.pt
22. RSYNC [Em linha]. [Consult. 29 Fev. 2004]. URL: http://samba.anu.edu.au/rsync
23. UNISON [Em linha]. [Consult. 29 Fev. 2004]. URL: http://www.cis.upenn.edu/~bcpierce/Unison/
24. DLIB [Em linha]. [Consult. 29 Fev. 2004]. URL: http://www.dlib.org/
25. Gutenberg [Em linha]. [Consult. 29 Fev. 2004]. URL: http://promo.net/pg/
26. Disputatio [Em linha]. [Consult. 29 Fev. 2004]. URL: http://www.disputatio.com/
27. Partidos Políticos [Em linha]. [Consult. 29 Fev. 2004]. URL: http://pesquisa.bn.pt/PartidosPoliticos/
28. Blogs em PT [Em linha]. [Consult. 29 Fev. 2004]. URL: http://blogsempt.blogspot.com/