PADRÕES PARA BIBLIOTECAS DIGITAIS ABERTAS E INTEROPERÁVEIS
STANDARDS TO OPEN AND INTEROPERABLE DIGITAL LIBRARIES

Luís Fernando Sayão
Comissão Nacional de Energia Nuclear, Centro de Informações Nucleares
lsayao@cnen.gov.br

Comente este artigo no blog Ebibli = http://encontros-bibli-blog.blogspot.com/

RESUMO
A interoperabilidade é uma das principais preocupações no desenvolvimento de sistemas de bibliotecas digitais abertas e que operam em rede. Entretanto, a interoperabilidade como forma de viabilização de intercâmbio de informações e de serviços cooperativos, exige a aplicação de um elenco de padrões abertos que cubram todos os processos de um repositório digital. A idéia deste documento é revisar os principais padrões, normas e protocolos que formam a infra-estrutura de uma biblioteca digital aberta e plenamente interoperável.

Palavras-chave: Bibliotecas digitais. Interoperabilidade. Normas. Padrões. Protocolos

1 INTRODUÇÃO

No momento em que as bibliotecas digitais se tornam, em escala mundial, um foco de intensa atividade de pesquisa e desenvolvimento, e se tornam também uma área de amplo interesse econômico e comercial, a todo instante se agregam novas questões e novas soluções, que lado a lado com os conceitos tradicionais da Ciência da Informação e da Biblioteconomia - revisitados e ampliados -, dão margem ao surgimento de um instigante e diversificado campo de estudo, envolvendo pessoas, recursos informacionais e tecnologia.

As bibliotecas digitais surgiram na década de 1990 e experimentam nos últimos anos um rápido crescimento que se irradia por todas as facetas que a área circunscreve: projeto, implementação, desenvolvimento e avaliação. As agendas de três importantes congressos realizados em 2002 – Joint Conference on Digital Libraries, 6th European Conference on Research and Advanced Technology for Digital Library e 5th International Conference on Asian Digital Libraries -, pela visão integrada e coerente das tendências da pesquisa e desenvolvimento na área de bibliotecas digitais que apresentaram, podem sintetizar com clareza o amplo espectro das vertentes de pesquisa nessa área (Shiri, 2003): a) arquiteturas, sistemas, ferramentas, e tecnologias b) conteúdos e coleções digitais; c) metadados; d) interoperabilidade; e) normas e padrões; f) organização do conhecimento; g) usuários e usabilidade; h) aspecto legal, organizacional, econômico e social. Não se pode esperar absolutamente que esses itens possam ser estudados de forma estanque; ao contrário, eles se mesclam conferindo aos objetos de pesquisa em biblioteca digital um alto grau de complexidade. Por exemplo, normas e padrões no contexto da pesquisa em bibliotecas digitais inclui todos os protocolos, regras e convenções que devem ser utilizados na arquitetura das bibliotecas

digitais, na formação das coleções, nos formatos, nas questões de interoperabilidade, na preservação digital, na atribuição de nomes permanentes, só para citar alguns itens.

O objetivo deste documento é fazer um levantamento sobre as principais regras que permitem o funcionamento das atividades que são pertinentes às bibliotecas digitais abertas, e que também criem a infra-estrutura necessária para viabilizar tecnicamente a interoperabilidade entre elas.

2 DEFINIÇÕES PRELIMINARES: PADRÕES, NORMAS, PROTOCOLOS E FORMATOS NO CONTEXTO DAS BIBLIOTECAS DIGITAIS

Uma biblioteca digital – no seu sentido pleno - não é meramente um repositório ou uma coleção de informações em formato digital; também não é somente uma tecnologia ou um conjunto de tecnologias que se pode avaliar isoladamente. Antes disso, é um sistema aberto, de múltiplas interligações e múltiplos subsistemas, envolvendo um ambiente organizacional, profissionais especializados provenientes de diversas áreas, recursos informacionais, usuários claramente definidos, tecnologia de informação, procedimentos, padrões e protocolos e, não menos importante, compromissos de longo prazo. A DLF - Digital Library Federation - torna patente essas inter-relações quando define bibliotecas digitais como:

> *"... organizações que proporcionam os recursos, incluindo pessoal especializado, para selecionar, estruturar, oferecer acesso intelectual, interpretar, distribuir, preservar a integridade e assegurar persistência através do tempo de coleções de objetos digitais, de forma que eles estejam prontamente disponíveis para uma comunidade alvo definida ou um conjunto de comunidades". (DLF, 2002)*

Neste contexto, as normas, padrões, formatos e protocolos cumprem um papel de fundamental importância, já que estabelecem as regras pelas quais os objetos são descritos, identificados e preservados, seus dados são armazenados, e os sistemas aos quais estão inseridos se comunicam. Considerando que o foco deste artigo está sobre as regras que viabilizam os processos das bibliotecas digitais abertas e interoperáveis, é melhor começar esclarecendo as diferenças entre elas, no contexto que mais nos interessa no momento.

a) **Padrões:** de acordo com a ISO – International Organization for Standardization - um padrão é um "documento aprovado por um organismo reconhecido que provê, pelo uso comum e repetitivo, regras, diretrizes ou características de produtos, processos ou serviços cuja obediência não é obrigatória" (ISO, 2006).
b) **Normas:** por outro lado, normas são as regras ou princípios estabelecidos sobre um determinado aspecto, que são definidas por quem de direito e não estão sujeitas à discussão. São os documentos técnicos que estabelecem padrões reguladores visando garantir a qualidade dos produtos, a racionalização da produção, o transporte e o consumo de bens, a segurança das pessoas e a uniformidade dos meios de expressão e comunicação.
c) **Protocolos:** são conjuntos de padrões contendo regras que governam as funções de comunicação num ambiente de rede. É realizado por meio da descrição do formato que a mensagem deve tomar e da maneira pela qual as mensagens são trocadas entre computadores (Haigh, 1998);
d) **Formato:** são as várias conformações na qual uma informação digital pode ser armazenada. O formato de um arquivo é um algoritmo expresso por um software para codificação de dados, bem como informações sobre o dado – estrutura, layout,

compressão. Existem centenas de formatos de arquivos, mas somente uns poucos são essenciais para as atividades de uma biblioteca digital (Haigh, 1998).

Existem os padrões proprietários, que são aqueles desenvolvidos ou promulgados por empresas privadas com o objetivo de assegurar ou ampliar as suas participações no mercado; e os padrões abertos, que são publicados e estão disponíveis para uso de quem quer que seja. Ambos os tipos podem se transformar em um padrão "de fato", isto é, um conjunto de regras ou especificações que são adotados largamente – e voluntariamente - pelo mercado e que se tornam normativos. Por outro lado, existem os padrões "de direito", que são aqueles endossados por organizações oficiais de normalização, em âmbito internacional, como a ISO (http://www.iso.org/) e a IETF (Internet Engineering Task Force - www.ietf.org/), ou em âmbito nacional como a NISO (National Information Standard Organization - www.niso.org/) nos Estados Unidos e ABNT (Associação Brasileira de Normas Técnicas - http://www.abnt.org.br/) no Brasil.

Entretanto, em todos os casos, o objetivo é tentar unificar a representação, manipulação ou a transmissão de algum item de informação de forma que dois ou mais diferentes sistemas possam "compreender" igualmente este item. Isto é a base do que se conhece por interoperabilidade entre sistemas de informação (Noerr, 2003).

3 OBJETOS DIGITAIS ADEQUADOS

A publicação da NISO "A Framework of Guidance for Building Good Digital Collections" (NISO, 2004a) nos oferece uma orientação acerca do que podemos considerar um "objeto digital adequado" (*good digital object*) inserido numa coleção digital. A arquitetura estabelecida pelo documento considera dois tipos de objetos digitais: objetos produzidos como representação ou substitutos de materiais em alguma forma analógica – livros impressos, manuscritos, peças de museus, etc. -, e objetos originalmente "nascidos digitais" – fotografias digitais, e-books, bases de dados científicas, imagens de exames médicos (tomografia, por exemplo), websistes, etc.

Um objeto pode estar completo em um único arquivo - por exemplo, um relatório em PDF -; ou pode consistir de múltiplos arquivos vinculados por *hiperlinks*, por exemplo, uma página HTML e as imagens vinculadas a ela; ou ainda consistir de múltiplos arquivos unificados por metadados estruturais, por exemplo, um livro digitalizado na forma de imagens de páginas separadas. "Nesse sentido, objetos digitais são conceitualmente equivalentes a itens que podem ser encontrados no acervo de bibliotecas, na coleção de museus, e em fundos arquivísticos" (NISO, 2004a, p.12).

Os seguintes princípios se aplicam aos objetos digitais adequados:

Princípio 1 – Prioridades da coleção, interoperabilidade e reusabilidade

Um objeto digital adequado deve ser produzido de forma que assegure a manutenção das prioridades da coleção ou coleções onde será inserido, enquanto mantém as qualidades que contribuam para a interoperabilidade e reusabilidade.

Princípio 2 – Persistência

Um objeto digital adequado deve ser persistente. Isto é, o objeto digital deve permanecer acessível através do tempo a despeito das mudanças tecnológicas.

Principio 3 – Padrões e melhores práticas no processo de digitalização

Um objeto digital deve ser digitalizado num formato que tenha como perspectiva o suporte aos usos atuais e os prováveis usos futuros, ou que possibilite a derivação de cópias de acesso que suportem esses usos. Como conseqüência, um objeto digital adequado deve ser permutável através de plataformas computacionais, deve ser

amplamente acessível e ser digitalizado de acordo com padrões reconhecidos e/ou melhores práticas. Quando não for possível a adoção de padrões e práticas, as razões para tal procedimento devem estar bem documentadas.

Princípio 4 – Identificadores únicos e persistentes

Um objeto digital adequado deverá ser designado por um identificador persistente e único que deverá estar em conformidade com esquemas de nomes bem documentados. Ele não deverá ser identificado tomando-se como referência nome de arquivos ou endereços de servidores – como o URL e outros endereços da Internet – cuja característica é a volatibilidade -, ao invés disso, identificadores estáveis devem ser capazes de ser resolvidos (mapeados) em endereços correntes.

Princípio 5 – Autenticidade, integridade, proveniência e contexto

Um objeto digital adequado pode ser autenticado no mínimo em três dimensões: primeira, um usuário deve ser capaz de determinar a origem, a estrutura e a história do desenvolvimento (proveniência e contexto); segunda, um usuário deve ser capaz de determinar se um objeto é o que diz ser (autenticidade); terceira, um usuário deve ser capaz de determinar se o objeto foi corrompido ou alterado de forma não autorizada e não documentada (integridade).

Princípio 6 – Metadados descritivos, administrativos e estruturais

Um objeto digital adequado deverá ter associado a ele um conjunto de metadados. Todo o objeto digital deverá ter metadados descritivos e administrativos. Os objetos digitais complexos – formados por vários arquivos – deverão estar associados a metadados estruturais.

A efetiva aplicação desses princípios é uma etapa importante na viabilização de repositórios digitais abertos e interoperáveis, mas para tal, as orientações estabelecidas devem ser traduzidas em melhores práticas e o uso de padrões abertos e de ampla aceitação. É precisamente sobre isso que será tratado a seguir.

4 PADRÕES DE FORMATOS PARA OBJETOS DIGITAIS

Via de regra existe uma correlação direta entre a qualidade de produção de um objeto digitalizado e a presteza e flexibilidade com as quais este objeto pode ser migrado através de plataformas computacionais. Dessa forma, a digitalização de objetos digitais, utilizando-se o mais alto grau de qualidade disponível, é compensada pela sua utilidade, versatilidade e maior acessibilidade a longo prazo. Entretanto, nem todos os objetos exigem este investimento, cada projeto de digitalização necessita determinar o valor dos objetos digitalizados e tomar as decisões apropriadas em relação à persistência e ao nível de interoperabilidade desejados (NISO, 2004a).

4.1 Imagens

No contexto dos repositórios digitais, a geração de imagens digitais tem dois propósitos primordiais: o primeiro deles é o armazenamento e a preservação; o outro, não menos importante, é o acesso. Esta dualidade de propósitos se expressa em exigências distintas - em termos de qualidade e de procedimentos de geração de imagens. A função básica de um arquivo de preservação é servir como um registro arquivístico de longo prazo e como fonte para a derivação de outros formatos de arquivos, através de edição ou conversão, mais apropriada para o acesso, a apresentação ou a transmissão via rede de computadores. Desta forma, objetos únicos, raros ou frágeis podem ser – dependendo do nível de acesso - compartilhados por um grande número de interessados, seja para a pesquisa, ensino ou qualquer outra finalidade

permitida pelo custodiante. (Princeton University Library). O arquivo de preservação é também uma descrição e/ou representação digital do objeto analógico original, na medida em que pode substituí-lo funcionalmente em termos de recuperação, apresentação e acesso. Limitando o acesso direto a documentos valiosos, ele pode ser usado também como um dispositivo de segurança prevenindo possíveis perdas ou danos que possa sofrer o objeto original causado por acidente, furto ou manuseio indevido. Para o caso de bibliotecas digitais, cujo foco principal é o conteúdo informacional - textual e iconográfico - dos objetos, e o objetivo primário do usuário é a recuperação e acesso, a representação digital é, na maioria das vezes, suficiente para os seus propósitos.

Arquivos digitais de preservação são criados como resultado direto da captura de uma imagem via processos de digitalização. Estes processos devem capturar o máximo de informação possível, posto que devem representar o mais precisamente possível a informação visual presente no objeto original. Devem ser gerados por meio de escanerização direta. Entretanto, se o objeto original não puder ser digitalizado diretamente – devido às suas dimensões ou por outro motivo -, pode ser necessário usar técnicas de fotografia como um passo intermediário. As fotografias intermediárias devem estar bem documentadas e representar o objeto original o mais precisamente possível. Há um consenso absoluto de que o processo de captura de imagem deve ser realizado de maneira correta e uma única vez. Isto é determinado para evitar manipulações desnecessárias e estresse físico sobre o material original, reduzindo dessa forma os riscos de deterioração e/ou a ocorrência de danos.

As imagens de uma coleção digital recaem nas seguintes categorias (Fleischhauer, 1998):

a) **Imagem de preservação**. Imagem de altíssima fidelidade usada para preservação, reprodução e derivação de outras imagens em formatos apropriados, tais como miniaturas e imagens de serviço. Quanto à compressão, a recomendação é que a imagem de preservação esteja livre de compressão; entretanto quando for imperativo o seu uso, ela deve ser sem perdas de informação e gerada por métodos não proprietários. É necessário enfatizar que a compressão adiciona um grau a mais de complexidade aos processos de migração voltados para a preservação de longo prazo (California Digital Library, 2001). Formato recomendado: TIFF ITU–T6(Tagged Image File Format) – formato de 24 bits, compressão interna, sem perda de dados (LZW), acionada opcionalmente. Profundidade de cor: escala de cinza de 8 bits, cor de 24 bits.
b) **Imagem de acesso ou de serviço**. Imagens de alta qualidade, comprimidas tendo em vista a otimização da recuperação, do acesso, da transmissão e da apresentação. Para cada registro a biblioteca pode oferecer um conjunto de imagens de acesso com graus variados de resolução. Formato recomendado: JPEG (Joint Photographic Expert Group) – formato de 24 bits, com o atributo de compressão com perda de dados, porém com alta taxa de compactação, apropriado para transmissão e apresentação, mas não para preservação. Profundidade de cor: escala de cinza de 8 bits, cor de 24 bits por pixel. Compressão de 10:1 para escala de cinza e 20:1 para cores
c) **Imagem prévia ou miniatura.** (*thumbnail image*). Uma imagem pequena apresentada para permitir ao usuário julgar se é de interesse recuperar a imagem de alta qualidade correspondente. Formato recomendado: GIF (Graphic Interchange

Format) – formato de 8 bits, compressão sem perdas de dados (LZW), apropriada para apresentações de baixa resolução. Profundidade de cor: cor de 8 bits por pixel.

4.2 Áudio

O fato dos arquivos de áudio serem relativamente volumosos tem implicado num empenho contínuo da indústria de multimídia na busca de novas e melhores soluções de compressão e também de esquemas de reprodução. Por esta razão, os formatos digitais de áudio apropriados para a distribuição via Web são menos estáveis do que os formatos digitais de texto e de imagem, e tendem a se tornar obsoletos muito rapidamente (Fleischhauer, 1998). Outra diferenciação importante dos arquivos de áudio diz respeito à impropriedade de se gerar arquivos de preservação. No caso de arquivos de imagem, quando uma nova versão de distribuição é necessária, o arquivo de preservação pode ser usado como fonte para a sua geração, já para o caso dos arquivos de áudio essa geração deve ser realizada a partir do item original ou de uma versão intermediária (Fleischhauer, 1998).

Os formatos de áudio digital, em termos de distribuição, são de dois tipos: *downloadable* e *streaming*. O primeiro tipo precisa ser "baixado" completamente para o computador do usuário antes de ser executado. Dado que esses arquivos são mais volumosos eles requerem tempo de transferência e espaço de armazenamento. O segundo tipo, também conhecido como áudio de fluxo contínuo, é executado à medida que está sendo transmitido através da rede e não pode ser gravado localmente. Esta tecnologia é que sustenta a transmissão de serviços de rádio via Internet, viabilizando a transmissão ao vivo ou serviços sob demanda que se tornarão importantes para as bibliotecas digitais, especialmente as voltadas para o ensino.

Os formatos para áudio digital mais comuns no domínio dos repositórios digitais são os seguintes (Noerr, 2003):

a) WAVE (extensão wav) – formado por seqüência de valores numéricos que são interpretados pelo computador, possui alta qualidade e facilidade de edição, porém resulta em arquivos volumosos impróprios para transmissão via Internet.
b) MPEG-1 Layer III (extensão mp3) – são arquivos semelhantes aos arquivos WAV, de elevada qualidade e extremamente compactados, e, portanto, adequados para transmissão por rede. O MP3 deu margem ao surgimento de uma forma inovadora de consumir música no mundo inteiro (HP do MPEG).
c) MIDI - sigla para Musical Instruments Digital Intarface – não é exatamente um formato de áudio, mas é uma interface digital para instrumentos digitais, constituindo-se numa especificação padronizada (protocolo) permitindo que instrumentos eletrônicos de qualquer fabricante se comuniquem uns com os outros e com o computador.
d) RealAudio (extensão .ra/rm) – formato proprietário da RealNetworks (http://www.realnetworks.com/) que inclui um avançado sistema de compressão e transmissão de áudio e video.

Além dos formatos, a qualidade do áudio digital depende ainda de outras especificações. De grande importância é a freqüência de amostragem (*sampling rate*), que indica a precisão com que um arquivo digital descreve o som analógico que ele representa; corresponde à freqüência que a medida de amplitude de um sinal analógico é tomada em um intervalo fixo de tempo no processo de conversão digital. Quanto maior

a freqüência de amostragem, melhor é a performance do arquivo de áudio. Por exemplo: um CD comercial tem a freqüência de amostragem de 44,1 kHz, significando que o sinal é medido 44.100 vezes por segundo. Devemos ainda considerar dados como o número de canais de gravação; se a gravação é mono, estéreo ou ainda quadrifônica, e dados sobre processamento de ruído. A decisão sobre a maioria dos parâmetros vai depender dos usos que se espera do registro.

4.3 Vídeo

Vídeo é na realidade uma seqüência de imagens acompanhadas, geralmente, de uma trilha sonora. Para o caso de vídeos digitais, a qualidade está associada à quantidade de quadros capturados por segundo (fps –*frame* por segundo) e da qualidade de cada quadro, que é expresso pela quantidade de *pixels* utilizado – que tem efeito no tamanho da tela - e pela profundidade de cor. É necessário também considerar a qualidade do vídeo original, para o caso de captura de fonte analógica, por exemplo, uma fita de videocassete, e a qualidade do áudio – com ou sem som, mono, estéreo, etc. A peculiaridade mais importante de um vídeo é que as imagens são ligeiramente diferentes das anteriores e das posteriores, indicando que uma forma especial de compressão pode ser usada tirando proveito dessa equivalência aproximada. Sem a compressão os arquivos serão extremamente volumosos. Por exemplo, um segundo de vídeo gravado à taxa de 30 fps, com dimensões de imagem de 640 x 480 pixels, com profundidade de cor de 24 bits requer aproximadamente 30 Mb para armazenamento (Noerr, 2003). Isto sinaliza a importância das técnicas de compressão para que se viabilize serviços de distribuição on-line de vídeos digitais.

Os formatos digitais para imagens em movimento adequados para a Web, como os de áudio, são também menos estáveis do que os formatos para texto e imagens, portanto estão mais suscetíveis à obsolescência tecnológica (Fleischhauer, 1998). Assim como para som digital, no ambiente Web, alguns provedores de conteúdo disponibilizam serviço de vídeo *streaming*, ou vídeo de fluxo contínuo. Essa tecnologia permite que o usuário possa assistir ao vídeo à medida que este é carregado no seu computador, sem a necessidade de "baixar" todo o arquivo para depois assisti-lo. Por este motivo, essa tecnologia é bastante usada na transmissão de eventos ao vivo e para a transmissão continua de radio e televisão via Internet. Esta modalidade apresenta algumas vantagens imediatas: permite uma melhor qualidade de reprodução; não necessita de muita largura de banda; e não permite que o arquivo seja copiado, como se pode fazer quando se utiliza o *downloading*. Os softwares necessários para o usuário executar esses tipos de arquivos são geralmente livres e estão disponíveis para a maioria dos sistemas operacionais.

É importante assinalar que recentes avanços nos processos de extração de imagem têm permitido que os vídeos digitais possam ser indexados por imagens selecionadas, e estas possam ser analisadas por suas características (Noerr, 2003). Na seção 7, que trata de metadados, isto é analisado com um grau a mais de detalhes.

A seguir são apresentados os padrões e formatos mais utilizados nos contexto das bibliotecas digitais.

a) AVI (extensão: avi) – sigla para Audio Video Interleaved -, significando entrelaçamento de áudio e vídeo. É um formato de arquivo audiovisual desenvolvido

pela Microsoft para a plataforma Windows, também utilizado para *streaming*. Roda no Windows Media Player.

b) MOV (extensão: mov) – formato de vídeo desenvolvido pela Apple para o seu programa multimídia, o Quick Time. Pode também ser usado para transmissão de fluxo contínuo (streaming).
c) Real Player (extensão: ra/rm) – formato proprietário da Real Network bastante utilizado para transmissões em *streaming* para áudio e vídeo.
d) MPEG (extensão: mpg/mpeg) – padrão para compactação de vídeo e áudio digitais.

O MPEG (Motion Pictures Expert Group) é um grupo de trabalho ligado à ISO/IEC, instalado em 1988, que tem como responsabilidade o desenvolvimento de padrões internacionais para compactação, descompactação, processamento e representação codificada de vídeo e áudio digitais e suas combinações, orientados para uma ampla faixa de aplicações. Como resultado de seu trabalho, o Grupo desenvolveu um conjunto de padrões voltados para compactação de vídeos/áudio identificados como o MPEG-1, o MPEG-2 e o MPEG-4. Esses padrões tornaram possível o vídeo interativo em CD-ROM, o DVD e a Televisão Digital. O MPEG-1 (ISO/IEC 11172), lançado em 1992, prevê a criação de vídeos com qualidade de videocassetes padrão VHS. Produtos tais como Vídeo CD e MP3 (MPEG-1 Audio Layer III) são baseados neste padrão. Por sua vez, o MPEG-2 (ISO/IEC 13818) – "Generic Coding of Moving Picture and Asociated Áudio" - é o padrão sobre o qual produtos como a televisão digital e o DVD estão baseados. O MPEG-4 (ISO/IEC 14496) – "Coding of audio-visual objects" – é o resultado de um esforço internacional envolvendo centenas de pesquisadores e engenheiros de todo o mundo na busca de um padrão para multimídia voltado para a Web fixa e a móvel. Ele foi desenhado para entregar vídeos com qualidade de DVD (MPEG-2) a baixas taxas de transmissão e arquivos menores. O MPEG-4 estabelece elementos tecnológicos padronizados que permitem a integração da produção, da distribuição e do acesso a conteúdos nas áreas da televisão digital, de aplicações gráficas interativas e de multimídia interativa. O padrão segue o paradigma de orientação a objetos e permite a codificação de objetos individuais. Um objeto pode ser um vídeo ou uma imagem (que podem ser sintéticos ou naturais), por exemplo, um pássaro em movimento ou uma fotografia de uma pessoa; pode ser ainda um objeto de áudio como o som de um instrumento de uma orquestra. A associação de um áudio e um vídeo é chamado de objeto áudio-visual. O padrão provê suporte para a representação de informação em 3D variável no tempo. Uma infra-estrutura para a gerencia e proteção de direitos relativos a objetos individuais também foi prevista pelo padrão (MPEG).

4.4 Textos Digitais

São arquivos textuais legíveis por máquina, produzidos com o apoio de editores de texto ou de outros programas similares, ou são resultados de processos de reconhecimento ótico de caracteres (OCR – Optical Caractere Recognition) a partir de imagens digitalizadas de textos impressos. No contexto de um repositório digital duas coisas devem acontecer a um arquivo textual: ele tem que ser armazenado de forma que possa ser exibido para o usuário quando requisitado; e precisa também ser processado e indexado permitindo que seu conteúdo esteja disponível para pesquisa (Noerr, 2003). Um problema importante vinculado aos arquivos textuais que deve ser superado é a codificação de caracteres.

A codificação de caracteres é a atribuição de um código computacional para cada letra do documento. Isto é feito durante a criação do documento, seja via processador de texto, ou por conversão por OCR da imagem escanerizada. Se os documentos provêm de fontes distintas e se, particularmente, são escritos em idiomas diferentes, é provável que eles estejam usando esquemas de codificação diferentes, e como conseqüência apareça algo sem sentido para os usuários e para os programas de indexação. Entretanto, existe um esquema único de codificação universal padronizado internacionalmente pela ISO, o Unicode - ou norma ISO 10646. Esse esquema foi desenvolvido pela Unicode Consortium (http://unicode.org/) e tem como compromisso "fornecer um único número para cada caracter, não importa a plataforma, não importa o programa, não importa a língua" (Unicode, 2005). Isto permite que quase todos as escritas atualmente em uso no mundo possam ser codificadas sem ambigüidades, além das escritas históricas já extintas e os símbolos, em especial os matemáticos e os musicais.

4.5 Reprodução Textual de Imagens e Textos Pesquisáveis

Em muitas situações, principalmente no mundo acadêmico, somente uma imagem não é o suficiente para o desenvolvimento de algumas atividades de pesquisa. Transcrições pesquisáveis de textos literários, livros raros ou manuscritos, por exemplo, podem ser de dramática importância para quem está procurando por padrões lingüísticos - palavras ou expressões particulares - dentro de um texto. Textos transcritos, especialmente quando codificados por intermédio de linguagem de marcação, podem facilitar imensamente o trabalho de navegação de pesquisadores (Fleischhauer, 1998)

A TEI, Iniciativa de Codificação Textual – Text Encoding Initiative (http://www.tei-c.org/) - é uma norma internacional e interdisciplinar, baseada em SGML, e mais recentemente também em XML (conhecida como TEI Lite), que capacita as bibliotecas, museus, editores e acadêmicos a representar todo o tipo de texto literário e lingüístico destinado à pesquisa on-line, ao ensino e à preservação, segundo uma DTD (Document Type Definition). O TEI Guidelines for Electronic Text Encoding and Interchange (http://www.tei-c.org/Guidelines2/) além de especificar como codificar os textos, dedica uma parte à especificação de um cabeçalho, que deve ser inserido no recurso, consistindo de metadados sobre a obra.

4.6 Aparência do Documento

A SGML, e de forma similar a XML, enquanto linguagens de marcação, têm como foco a descrição do conteúdo e da estrutura do documento; porém, em muitos casos, é necessário também descrever a aparência e o layout do documento e apresentá-lo em forma impressa e na tela de um equipamento. Para estes casos existe um consenso em torno do PDF – Portable Document Interface. O PDF é um formato proprietário da empresa Adobe (http://www.adobe.com/), e faz parte de uma suíte de produtos de software – o Acrobat - destinada à criação, edição, exibição, etc. de documentos eletrônicos. Apesar de ser um formato proprietário, sua especificação encontra-se aberta e disponível. Entretanto, o PDF não considera aspectos de preservação de longo prazo. Para contornar este problema, estabeleceu-se em 2002 uma iniciativa conjunta, envolvendo instituições de diferentes origens para criar um formato padrão, baseado no PDF, para documentos arquivados eletronicamente. Como resultado desta iniciativa, em setembro de 2005 a ISO aprovou uma nova norma, a ISO1905-1 – Document Management – Electronic file format for long-term preservation – Part 1: use of PDF

1.4 (PDF/A-1) que define "um formato de arquivo baseado em PDF, conhecido como PDF/A, que provê um mecanismo para representar documentos eletrônicos de forma que seja preservada a aparência visual destes ao longo do tempo, independente de ferramentas e sistemas usados para criação, armazenamento ou apresentação do arquivo" (PDF Tools AG, 2006, p.5). O PDF/A guarda as seguintes características: independente de plataforma de hardware e software; autocontido, significando que contém todos os recursos necessários – inclusive fontes – para exibição/impressão; autodocumentado, contém as suas próprias descrições embutidas nos arquivos PDF via Adobe Extensible Metadata Platform (XMP); sem criptografia, ausência de mecanismos de proteção de arquivos; e baseado numa especificação oficial disponível publicamente (AIIM, 2006).

4.7 Textos formatados, apresentações, planilhas, etc.

Observa-se uma tendência acentuada, principalmente por parte das agências governamentais, de se estabelecer políticas de interoperabilidade que adotem padrões abertos para os aplicativos voltados para escritório – editor de texto, planilhas eletrônicas, apresentações de slides, banco de dados, manipulação de imagens, etc. –, hoje dominado fortemente pelo pacote de programas MS Office, de propriedade da Microsoft. A escolha tem recaído sobre o Open Document Format (ODF), um conjunto de formatos de arquivos para aplicação de escritório desenvolvido para estabelecer um padrão de mercado, cuja homologação pela ISO aconteceu em maio de 2006 (ISO 26300). Por se tratar de um padrão aberto, qualquer software pode implementá-lo, tornando mais fácil a manipulação, distribuição e preservação de documentos criados sob este padrão.

O ODF partiu de uma iniciativa da OASIS – Organization for the Advancement of Structured Information Standard – um consórcio internacional criado com o objetivo de desenvolver e promover padrões para formatos digitais, especialmente para utilização na Web. A base do ODF é um esquema XML inicialmente criado pelos desenvolvedores do Open Office.org (Alecrim, 2006)

5 IDENTIFICADORES PERMANENTES

O sucesso de um sistema de informações distribuídas – tal como é caracterizada a própria Web – depende fortemente da vinculação consistente entre os recursos que estão disponibilizados *on-line*. Isto se concretiza por intermédio da estabilidade dos *links* que estão presentes nos catálogos, nos índices e nas listas que constituem os diversos serviços de descoberta de recursos. Contudo, para que isso se efetive é necessário que para cada recurso seja atribuído um nome que o identifique permanentemente, sem ambigüidades e independente de localização. A vinculação entre recursos pode variar de uma simples inserção de um *link* com o sentido de enriquecer uma informação - por exemplo, um *hiperlink* no nome de uma organização que é mencionada num documento -, até a citação formal a um outro recurso, expresso por uma referência bibliográfica. No domínio das bibliotecas digitais, a categoria de *links* que representa uma citação formal, isto é, a que formaliza uma vinculação de um recurso a outro, é a mais significativa para a sustentação da integridade do conteúdo intelectual, constituindo o que a literatura chama de "*link* referencial" (Caplan; Arms,1999). Para o desenvolvimento pleno de bibliotecas e arquivos digitais de arquitetura aberta e plenamente interoperáveis, há um interesse contínuo por parte dos vários atores envolvidos, que incluem principalmente as organizações responsáveis pelo

ordenamento da Internet, as organizações internacionais de normalização e as organizações voltadas para o desenvolvimento de arquivos e bibliotecas digitais, de superarem a fragilidade dos esquemas atuais de identificação de recursos na Web, através do estabelecimento de mecanismos persistentes de identificação que sejam mais robustos e independentes de fatores voláteis, como os vinculados ao endereço do recurso na Web e/ou nas tecnologias e protocolos de uso corrente (Payette, 1998).

5.1 O problema de Identificação de Recursos na Internet – Nome versus Localização

> "Nomes são blocos de estrutura vitais para as bibliotecas digitais. Eles são necessários para identificar objetos digitais, para registrar propriedade intelectual vinculada a esses objetos e para registrar mudanças na propriedade dos objetos digitais. Eles são necessários para citação, para recuperação de informação, e são usados como *links* entre objetos" (Arms, 1995).

Para desempenhar este papel, os nomes têm que ser únicos e persistentes. Um nome persistente no contexto dos repositórios digitais é compreendido como um identificador único que perdure por um período tão longo quanto seja necessário; que perdure mesmo que a organização que atribuiu o nome ao objeto não mais exista quando o objeto for usado. Para isto funcionar é necessário estabelecer uma infra-estrutura administrativa para decidir quem pode assinalar nomes que identifiquem univocamente os recursos digitais de forma persistente. É imprescindível também criar sistemas automatizados - sistemas de resolução - capacitados a resolver nomes rapidamente, ou seja, sistemas que dêem como resposta o endereço onde está armazenado um objeto digital detentor de um dado nome (Arms, 1995).

Nos dias de hoje, o URL - Uniform Resource Locator - é a porta de entrada para os conteúdos que estão disponíveis na Web. Ele define, como o seu próprio nome diz, a localização do recurso. De maneira diferente do número de chamada de um livro, que é raramente reclassificado, o maior problema com os URL´s é que eles mudam com muita freqüência, esta é precisamente a maior fragilidade desse esquema. A crítica mais contundente que se faz ao URL é que ele mistura numa mesma cadeia de caracteres elementos que desejavelmente deveriam estar separados, como o método pelo qual o recurso é acessado, o nome da máquina servidora e o caminho (*path*) do documento, isto é, sua localização, e o nome do arquivo, que nem sempre é único (Cleveland, 1998). Apesar dessa fragilidade, freqüentemente tratamos o URL como se ele fosse um identificador formal para representar um objeto digital. Na realidade, o URL é simplesmente um endereço mascarado como um identificador. Confiar nele como um identificador único para os recursos digitais, é como usar o endereço residencial de uma pessoa no lugar do seu CPF (Payette, 1998). O efeito da instabilidade do URL está continuamente presente para quem busca profissionalmente informação ou simplesmente navega na Web. Repetidamente estamos diante de links quebrados, e a cada dia cresce a proporção de links que produzem como resposta ao seu acionamento a mensagem recorrente: "Erro HTTP - 404 Arquivo não encontrado". Estima-se que esse erro ocorre em cerca de 7% dos *links* (Bigwood, 1999). Numa medida ao longo do tempo, tendo como objeto de estudo uma amostra aleatoriamente selecionada de URL´s, foi demonstrado que somente cerca de 34% dos URL´s permaneciam ativos depois de um período de quatro anos (Koehler, 2002).

5.2 Esquemas de Identificação Permanente

5.2.1 URN – Uniform Resource Name

O esquema de identificadores usados atualmente na Internet é o URI - Uniform Resource Identifiers -, cuja especificação está descrita no documento publicado em 1988 "RFC 2396 – Uniform Resouce Identifiers (URI): Generic Syntax" (Berners-Lee et al, 1988). Nesse documento se distinguem duas grandes categorias de URI´s: URL (Uniform Resource Locator) e o URN (Uniform Resource Name). O URL "se refere ao subconjunto do URI que identifica o recurso através da representação do seu mecanismo primário de acesso (por exemplo, sua localização na rede)"; enquanto o URN "se refere ao subconjunto do URI que é preciso permanecer globalmente único e persistente mesmo quando o recurso deixa de existir ou se torna indisponível" (Berners-Lee et al, 1988). Dessas duas categorias, a única que está plenamente desenvolvida e disseminada é o URL (Dack, 2001). A RFC 1737- Functional Requirements for Uniform Resources Names, fixou a concordância geral em relação aos requisitos básicos para URN´s e, por definição, para qualquer outro esquema útil de identificadores persistentes. Esses requisitos são os seguintes (Dack, 2001; Sollins; Masinter, 1994):

a) **Escopo global**: o URN é um nome com validade em escala global, o que não implica em dependência relativa à localização. Ele tem o mesmo significado em qualquer lugar.
b) **Unicidade global**: o mesmo URN não deverá nunca ser assinalado para dois recursos diferentes;
c) **Persistência**: o tempo de existência de um URN deve ser permanente. Isto é, o URN será globalmente único para sempre, e pode ser usado como referência para um recurso digital além da existência do recurso e da organização que lhe atribui o nome;
d) **Escalabilidade**: URN pode ser assinalado para qualquer recurso que venha a ser concebido para disponibilização na rede por centenas de anos;
e) **Suporte aos sistemas legados**: o esquema deve permitir a incorporação dos sistemas de identificação já existentes – sistemas legados - que satisfaçam as demais exigências aqui descritas. Por exemplo, códigos ISBN e ISSN, identificadores ISO e outros de igual importância;
f) **Extensibilidade**: qualquer esquema de URN´s deve permitir extensões futuras para o esquema;
g) **Independência**: as condições sob as quais um nome pode ser atribuído são de responsabilidade única e exclusiva da autoridade que atribui nomes**;**
h) **Resolução**: o URN não impede a resolução (tradução para um URL). Mais especificamente, para URN´s que têm URL´s correspondentes, deve existir algum mecanismo viável para traduzir um URN em um URL.

Para o pleno funcionamento do esquema URN seria necessário, entre outras coisas, que fosse incorporado aos navegadores Web padrões – Internet Explorer, Netscape, etc. – a capacidade de reconhecer e direcionar os URN´s. A IETF continua a trabalhar em conjunto com toda a comunidade Internet, incluídos os desenvolvedores de navegadores Web, no intuito de alcançar consenso sobre os padrões necessários para tal. Enquanto isso não acontece, todos os esquemas de URN já implementados, e os que o

serão a curto e médio prazo terão que utilizar uma intermediação, ou seja, terão que lançar mão de servidores *proxy* que direcionam a requisição para o computador servidor onde está localizado o recurso que se deseja. Isto significa que os identificadores para funcionarem têm que estar encapsulados em URL's que os direcionem ao endereço do servidor *proxy*. Isto será visto em alguns exemplos de implementações de identificadores persistentes que analisaremos a seguir: Persistent URL(PURL), Handle System e Digital Object Identifier (DOI).

5.2.2 PURL – Persistent URL

O PURL (http://purl.oclc.org/) é um esquema desenvolvido pela OCLC – On-line Computer Library Center (http://www.oclc.org/) - na tentativa de separar o nome de um recurso Internet de sua localização, e dessa forma, aumentar a probabilidade de que ele esteja disponível quando o seu link for acionado. Um dos objetivos subjacentes ao esquema PURL é contornar a atual falta de consenso e de progresso nas questões de nomes na Internet e, ao mesmo tempo, estabelecer práticas concernentes ao uso de identificadores persistentes em sistemas bibliográficos. Nessa direção, o sistema PURL preenche algumas das especificações do URN, e pode ser considerado uma solução intermediária aguardando o momento em que o URN se torne uma parte efetiva da arquitetura de informação da Web. A OCLC tem como perspectiva, para um futuro oportuno, poder traduzir mecanicamente o sistema PURL para a forma URN (Weibel; Jul; Shafer, 1995).

O PURL é funcionalmente um URL cujo endereço do servidor é o nome do serviço de resolução PURL. Ou seja, ao invés de apontar diretamente para a localização do recurso, o PURL aponta para um serviço de resolução intermediário, que por sua vez consulta uma base de dados que armazena a localização atualizada do recurso. Em termos de tecnologia web, trata-se simplesmente de um redirecionamento padrão HTTP – Hypertext Transfer Protocol - no qual não se exige uso de outros protocolos e não há exigência de modificações no software cliente (Shafer et al, 1996). Dessa forma, o PURL minimiza o problema da transitoriedade da localização do URL na medida em que ele é sempre traduzido pelo URL corrente do recurso. Uma vez que o PURL de um recurso é registrado na base de dados, ele pode ser usado para substituir o URL atual do objeto em diversas instâncias onde é desejável maior estabilidade, tais como registros bibliográficos em catálogos, ou em outros tipos de metadados que estabelecem *link* com o recurso. Nesta situação, se o URL do recurso mudar por algum motivo, a única ação de manutenção necessária é a atualização na base de dados do serviço de resolução. O Projeto InterCat demonstrou como o PURL pode ser usado em catálogos on-line, incluindo os PURL´s no campo 856 do MARC (Weibel; Jul; Shafer, 1995; Payette, 1998).

5.2.3 Handle System

O Handle System (http://www.handle.net) é um sistema distribuído de computadores concebido para assinalar, armazenar, administrar e resolver identificadores ou nomes persistentes de objetos digitais conhecidos como *handles.* Esses nomes podem ser rapidamente resolvidos em informação necessária para localizar e acessar os objetos a que se referem num ambiente de rede, tal como a Internet. O Handle System é caracterizado como um sistema de informação de grande amplitude, projetado para alcançar interoperabilidade global através de uma rede hierarquicamente distribuída de servidores. O objetivo inicial do Sistema era estabelecer uma infra-

estrutura básica no domínio das bibliotecas digitais e das publicações eletrônicas; entretanto, o Handle System criou uma arquitetura de tal forma abrangente que aplicações de propósitos diversos podem ser hoje desenvolvidas tomando-a como base (Dack, 2001; Payette, 1998). O Sistema constitui-se também numa implementação do URN – Uniform Resource Names -, posto que seu esquema de nomes, resolução e estrutura de registros estão em conformidade com as especificações RFC 1737, "Functional Requeriments for Uniform Resource Names" (Sollins; Mainter ,1994).

O Handle System foi desenvolvido pelo CNRI - Corporate for National Research Initiatives (http://www.cnri.reston.va.us/) - e teve sua origem no projeto NCSTRL -Networked Computer Science Technical Report Library (http://www.ncstrl.org/). Parte desse projeto estava voltada para o estabelecimento de uma arquitetura para apoiar o desenvolvimento de uma biblioteca digital aberta e distribuída (Dack, 2001). Uma aplicação do Handle System relativamente conhecida no Brasil é a adotada pelo software Dspace (http://www.dspace.org/). Entretanto, o exemplo mais conhecido em escala global é o DOI System, que é uma aplicação do Handle System voltada para gestão de *copyright* de objetos digitais. Este sistema será visto a seguir.

5.2.4 DOI – Digital Object Identifier

O Digital Object Identifier (http://www.doi.org) é uma aplicação específica do Handle System voltada para a identificação persistente de recursos digitais sobre os quais possam ser atribuídos direitos de propriedade intelectual, bem como para o intercâmbio de informações sobre essas propriedades em um ambiente de rede. Além de identificar, o DOI associa aos objetos digitais dados estruturados, ou seja metadados - informações bibliográficas e comerciais passíveis de atualização. Nessa direção, o DOI estabelece uma infra-estrutura ampla, que tem como perspectiva ligar os usuários aos fornecedores de conteúdo, dentro de um escopo que considera sempre a facilitação das práticas de comércio eletrônico de conteúdos e a viabilidade da gestão automática de *copyright*. Dessa forma, o escopo do DOI se estende além da intenção de ser um simples identificador, abrangendo também os desafios de gerenciar o comércio eletrônico, as questões ligadas à gerência de direitos para objetos digitais publicados na Internet, o controle de transações e ainda a comunicação entre os editores e seus clientes (Payette, 1998). O DOI, enquanto uma aplicação do Handle System, se constitui também numa implementação dos conceitos circunscritos pelo *Uniform Resource Name* (URN) e pelo *Uniform Resouce Identifier* (URI), no entanto a metodologia DOI apresenta um diferencial importante: ela agrega uma infra-estrutura técnica e social a estes conceitos teóricos, que é viabilizada pela adoção de um Modelo de Dados.

O DOI foi oficialmente lançado na Frankfurt Book Fair em 1997. O seu desenvolvimento é de responsabilidade da International DOI Foundation (IDF), uma instituição sem fins lucrativos fundada por iniciativa da Association of American Publisher (AAP), cujo corpo de membros é composto por representantes de grandes editores, publicadores e empresas de software, bem como organizações que representam os interesses dos detentores de direitos autorais, como por exemplo, a International Publishers Association (Dack, 2001).

6 PRESERVAÇÃO DIGITAL

As bibliotecas digitais hoje são geradoras, custodiantes e distribuidoras de materiais informacionais digitais de toda natureza, e como tal têm a responsabilidade inicial de assegurar a preservação de longo prazo e, conseqüentemente, a plena

acessibilidade a esses materiais. Há um consenso claro de que, na medida em que os criadores, disseminadores e proprietários de informação digital aceitam a responsabilidade inicial de arquivarem seus objetos digitais, terão também que incorporar preservação digital nos seus processos. Dessa forma, as bibliotecas digitais se tornam cada vez mais um elo importante na perenização dos estoques de informação digital, os quais constituem testemunhos das atividades da organização ou sistema de organizações no qual essas bibliotecas estão inseridas. Um exemplo concreto disso são as redes de bibliotecas de teses e dissertações, cujos acervos digitais distribuídos em rede são o testemunho das atividades acadêmicas e de pesquisa de uma universidade, de um país ou de uma região (Marcondes; Sayão, 2003a).

"A preservação digital envolve não somente a retenção do objeto informacional em si, mas também do seu significado. É necessário, portanto, que as técnicas de preservação sejam capazes de compreender e recriar a forma original ou a função do objeto de forma que sejam asseguradas sua autenticidade e acessibilidade". Conclui-se, portanto, que para manter os objetos digitais perenemente acessíveis para uso, se requer algo mais do que preservar simplesmente o artefato físico; é necessário considerar também várias outras dimensões que o problema apresenta: 1) a preservação física, cujo foco está na preservação das mídias e na sua renovação quando se fizer necessário; 2) a preservação lógica, que tem como foco os formatos e a dependência de hardware e software que mantenha legíveis e interpretáveis a cadeia de bits; 3) preservação intelectual, que tem como foco o conteúdo intelectual e sua autenticidade e integridade; 4) é importante ainda considerar a preservação do aparato – na forma de metadados - necessária para localizar, recuperar e representar a informação digital; 5) assim como proceder ao monitoramento e à instrumentalização da comunidade alvo, audiência para o qual a informação de forma privilegiada se dirige, no sentido de garantir que ele possa compreender plenamente a informação no momento do seu acesso (Sayão, 2006).

Dois documentos têm importância fundamental na formalização e na conceitualização do problema da preservação dos objetos digitais e também no estabelecimento de um elenco de informações necessárias para instruir os processos de preservação. São eles: o relatório Preserving Digital Information (Task Force on Archiving Digital Information, 1996) e o modelo de referência OAIS/ISO - Open Archival Information System (CCSDS, 2002) . O Preserving Digital Information, foi produzido pelo Task Force on Archiving of Digital Information em 1966, no âmbito da Commission on Preservation and Access (CPA) e do Research Libraries Group (RLG); enquanto o modelo de referência OAIS foi publicado em 2003 pela ISO e teve o desenvolvimento coordenado pelo Consultative Committe for Space Data System, de sigla CCSDS (http://public.ccsds.org/default.aspx), ligado à NASA – National Aeronautics and Space Administration. O OAIS tinha como objetivo estabelecer padrões para dar apoio à preservação de longo prazo de informações digitais decorrentes de observações espaciais. Embora o desenvolvimento do modelo tenha se originado no âmbito da comunidade espacial, ele se dirige também a outras audiências. O documento define um modelo de referência de alto nível para um sistema aberto voltado para arquivamento de informações. No Modelo o sistema de arquivamento é definido como "uma organização de pessoas e sistemas que assume a responsabilidade de preservar informação e torná-la disponível – em termos de acesso – para uma classe privilegiada de usuários, designada como Comunidade Alvo". Entre outras coisas, o Modelo OAIS tem o objetivo de proporcionar uma arquitetura comum, que pode ser usada para ajudar a compreender o desafio das organizações que têm a responsabilidade de arquivo,

especialmente aquelas envolvidas com a informação digital e a sua preservação (Day, 1999; Lavoie, 2000).

6.1 As Estratégias de Preservação Digital

Até o presente momento - e provavelmente até um futuro indefinido - não teremos disponível uma estratégia única capaz de dar conta de todo o espectro de problemas relacionados à preservação digital. O que se apresenta são soluções específicas para casos específicos. Dentre as soluções potencialmente completas, algumas são de uso corrente, mesmo que em pequena escala, outras são experimentais e estão ainda nas bancadas dos laboratórios ou ainda em patamares bastante teóricos (Lee et al, 2002). De forma geral, a comunidade envolvida no problema de preservação digital foi capaz de desenvolver métodos efetivos para a preservação do conteúdo informacional de materiais digitais, cujos formatos e estruturas são bem conhecidos e mais simples, mas falha ou obtém resultados pífios para materiais mais complexos e dinâmicos, ou para os que constituem formas novas ou emergentes de documentos e que expressam a inovação no uso da tecnologia digital.

Um fator fundamental de sucesso para o aumento da longevidade dos objetos digitais, não importando a estratégia de preservação digital adotada, está relacionado com a adoção de padrões, especialmente os padrões abertos. Numa visão otimista, os padrões abertos permitem que os documentos digitais sejam representados em formatos mais duradouros e estáveis, dessa forma reduzindo a velocidade do ciclo de obsolescência. A aplicação de padrões na preservação digital – na codificação, nos formatos e nos esquemas de representação - torna os processos de preservação digital mais fáceis, menos freqüentes e mais baratos, na medida em que reduzem a grande variedade de processos de preservação customizados, que são decorrentes da multiplicidade de formatos em que se traduzem os objetos digitais não padronizados. Idealmente, a padronização deve preceder a própria criação do objeto da preservação (Hedstrom, 1997/1998). Existe um consenso claro entre os especialistas da área de que a preservação digital pode ser realmente facilitada através da adoção de alguns procedimentos que incluem: 1) a definição de um conjunto limitado e gerenciável de padrões, preferencialmente abertos e/ou de ampla aceitação e de uso corrente; 2) a aplicação desses padrões na criação de novos objetos digitais, ou na conversão de documentos analógicos para formatos digitais; 3) o acompanhamento da obsolescência dos padrões desse conjunto e o monitoramento do surgimento de novos padrões; 4) a migração para os novos padrões tão logo eles estejam consolidados (Bullock, 1999).

Outro fator de fundamental importância para a gestão da preservação digital é a adoção de esquemas de metadados que registrem informações necessárias para a preservação. O OAIS estabelece como essenciais as seguintes informações: 1) a referência – registra identificação do objeto informacional no contexto local e global e ainda a sua descrição; 2) o contexto - documenta a relação do objeto digital com seu ambiente, incluindo a razão de sua criação e como ele está vinculado com outros objetos do universo digital e as dependências tecnológicas de hardware e software; 3)

a proveniência - refere-se à informação que documenta a cronologia do conteúdo informacional, incluindo sua origem ou fonte, as mudanças do seu conteúdo sofridas ao longo do tempo, a cadeia de custódia, propriedade intelectual, permissões de acesso, etc.; 4) a permanência - documenta os mecanismos de autenticação usados para assegurar que o conteúdo informacional de um objeto digital não sofreu alterações não documentadas

, como assinatura digital, certificação digital, criptografia, etc. (Sayão, 2006)

7 METADADOS

7.1 Definições

A NISO entende metadados como "informação estruturada que descreve, explica, localiza, ou ainda possibilita que um recurso informacional seja fácil de recuperar, usar ou gerenciar. O termo metadados freqüentemente designa dados sobre dados, ou informação sobre informação" (NISO, 2004b, p.1). As funções dos metadados compreendem a descoberta de recursos - que permitem que recursos sejam identificados, localizados, selecionados por critérios de relevância e distinguidos por diferenças e similaridades; a organização de recursos; a facilitação da interoperabilidade; a identificação digital; e a preservação digital. Existem basicamente três tipos de metadados (NISO, 2004b):

a) **Metadados descritivos**: são os metadados que descrevem um recurso com o propósito de descoberta e identificação. Eles podem incluir elementos tais como título, autor, resumo e palavras-chave;
b) **Metadados estruturais**: são os metadados que indicam como objetos compostos por vários elementos são recompostos. Por exemplo, como as páginas de um livro, digitalizadas separadamente, são ordenadas para formar um capítulo;
c) **Metadados administrativos**: fornecem informações que ajudam no gerenciamento de um recurso informacional. Por exemplo: informações sobre como e quando o recurso foi criado, informações técnicas sobre o arquivo e sobre quem possui direito de acesso a ele. Existem vários subconjuntos de dados administrativos, mas dois deles têm sempre um destaque especial e são listados muitas vezes como tipos específicos de metadados, são eles: metadados para gerenciamento de direitos: fornecem informações sobre os direitos de propriedade intelectual relacionados a um determinado recurso; e metadados para preservação: guardam informações necessárias a gestão da preservação digital de um determinado recurso (ver item 6.1).

Metadados podem descrever recursos informacionais em qualquer nível de agregação – uma coleção, um recurso simples, ou um elemento que faz parte de um outro recurso, como, por exemplo, uma fotografia inserida num artigo; pode ainda ser usado para descrever um trabalho, bem como uma manifestação ou uma expressão desse trabalho, por exemplo, um relatório, uma edição particular desse relatório, ou uma cópia específica da edição. Os metadados podem estar embutidos num objeto digital – inscritos na sua codificação, como é comum nos documentos HTML; ou podem estar armazenados separadamente, estruturados em base de dados e vinculados ao objeto que eles descrevem, facilitando a busca e a recuperação.

7.2 Esquemas de Metadados

Esquemas de metadados são conjuntos elementos de metadados projetados para um propósito especifico como, por exemplo, descrever um recurso informacional. A definição ou significado dos elementos é conhecido como a semântica do esquema, e os valores de um dado elemento são os conteúdos. Os esquemas de metadados, geralmente, especificam os nomes dos elementos e as semânticas correspondentes. A American Library Association (ALA) estabelece que um esquema de metadados fornece uma estrutura formal desenhada para identificar a estrutura do conhecimento de uma

determinada disciplina, e vincular essa estrutura à informação da disciplina por meio da criação de um sistema de informação, que auxiliará na identificação, descoberta e uso da informação no âmbito dessa mesma disciplina (ALA, 2000). A seguir são revisados os esquemas e estruturas de metadados mais significativos para a área de bibliotecas digitais.

7.2.1 Descrição bibliográfica

MARC – Machine-Readable Cataloguing (http://lcweb.loc.gov/marc/) - é um formato desenvolvido por iniciativa da Biblioteca do Congresso Americano iniciado há trinta anos. Os elementos de dados do MARC formam a base da maioria dos catálogos usados hoje em bibliotecas de todo mundo. O MARC passou a ser USMARC nos anos 80 e MARC 21 no final dos anos 90. A Library of Congress Network e o MARC Standards Office estão desenvolvendo uma infra-estrutura para trabalhar com dados MARC em ambiente XML, que disponibiliza um conjunto de componentes - tais como esquemas, folhas de estilo, ferramentas de softwares desenvolvidos e mantidos pela LC - que permite que os usuários trabalhem com dados MARC de acordo com as suas necessidades específicas (Library of Congress, 2004).

MODS – Metadata Object Description Schema (http://www.loc.gov/standards/mods/) - é um esquema de metadados descritivos, derivados do MARC 21, desenvolvido também pela Library of Congress. Ele é expresso na linguagem de esquema XML e é, em parte, uma reposta para a necessidade de se ter um formato bibliográfico de metadados que não seja específico para a comunidade de bibliotecas, mas que faça uma ponte entre aplicações bibliográficas tradicionais e aplicações bibliográficas que tenham outros propósitos. O esquema MODS caracteriza-se por poder incluir dados selecionados de registros MARC 21 já existentes – ou seja, traduzir registro MARC em XML -, bem como de ser capaz de gerar registros originais de descrição de recursos. Uma descrição rica de recursos digitais é um foco particular do MODS, que apresenta algumas vantagens sobre outros esquemas. Por exemplo: os elementos do MODS são mais completos do que o Dublin Core; são mais compatíveis com dados de bibliotecas do que o ONIX e o Dublin Core; e são mais simples de aplicar do que o MARC 21. Além disso, através do uso da linguagem de esquema XML, o MODS oferece uma série de aprimoramentos em relação ao MARC (NISO, 2004b).

DUBLIN CORE METADATA ELEMENT SET (http://dublincore.org/) é a língua franca para a representação de informações na Web.. Seu objetivo original era definir um conjunto de elementos que pudesse ser usado pelos próprios autores para descrever seus trabalhos publicados na Web, tendo em vista a proliferação de recursos eletrônicos e a impossibilidade dos profissionais de informação em catalogá-los. A idéia era definir uns poucos elementos e algumas regras simples. Originalmente o esquema contava com 13 elementos, hoje são 15. Esses elementos são núcleos de representação que viabilizam interoperabilidade entre vários sistemas Controvérsias e discussões levaram a alguns refinamentos semânticos e a ampliação para atender a interesses de comunidades específicas. Todos os elementos do DC são opcionais e repetitiveis e podem ser apresentados em qualquer ordem (NISO, 2004b).

7.2.2 Arquivos

EAD – Encoded Archival Description (http://www.loc.gov/ead/) - é um conjunto de regras, formando uma estrutura, desenvolvido como um meio para marcação de partes intelectuais e físicas de instrumentos de localização de informações

contidas em arquivos – tais como inventários, guias ou catálogos -, de forma que estas possam ser buscadas, recuperadas, exibidas e intercambiadas entre computadores e pessoas independente de plataforma (EAD, 2006). As regras do EAD são escritas na forma de uma DTD/SGML - Document Type Definition/Standard Generalized Markup Language-, que usa representações codificadas de elementos voltadas para o processamento eficiente por máquina (NISO, 2004b; Gómez; Pitti, 1999).

7.2.3 Objetos Educacionais

LOM – Learning Object Metadata (http://ltsc.ieee.org/wg12/) - é um modelo de dados, geralmente codificado em XML, usado para descrever adequadamente objetos de aprendizagem. Estabelecido pela norma IEEE 1484.12.2-2002, o LOM foi desenvolvido pelo IEEE Learning Technology Standards Commiteee (LTSC) tendo como foco a definição de um conjunto mínimo de atributos necessários ao gerenciamento, localização e avaliação de objetos de aprendizagem, que são definidos neste contexto como qualquer entidade, digital ou não, que pode ser usada, reusada durante o aprendizado apoiado por tecnologia - educação à distância e treinamento baseado em computador, ambiente interativo de aprendizagem, etc. Exemplos de objetos de aprendizagem incluem itens tais como conteúdo multimídia, ferramentas de software, softwares educacionais, bem como pessoas, organizações e eventos. Os atributos são agrupados em nove categorias (NISO, 2004b; Ogbuji, 2003): 1) Geral – contém informação sobre o objeto como um todo; 2) Ciclo de vida – contém metadados sobre a evolução do objeto; 3) Meta-metadados – informa sobre os metadados que descrevem o objeto; 4) Técnico – apresenta descrição de características e requisitos técnicos; 5) Educacional – contém atributos educacionais e pedagógicos; 6) Direitos – descreve direitos relacionados à propriedade intelectual e condições de uso; 7) Relação – identifica objetos relacionados; 8) Notação – contém comentários e a data, além do autor do comentário; 9) Classificação – descreve o objeto em relação a um sistema de classificação particular (IEEE).

7.2.4 Comércio Eletrônico

Os esquemas de metadados estão crescentemente sendo desenvolvidos para dar apoio ao comércio eletrônico e à gestão de direitos relacionados à propriedade intelectual. Nesta categoria se enquadram o <indecs> e o ONIX.

<indecs> Framework – Interoperability of Data in Commerce System - é fundamentalmente um modelo de dados para a propriedade intelectual e sua transferência. Ele é resultado de um esforço colaborativo internacional apoiado pelo European Commision´s Info 2000 Programme, cujos participantes são detentores de direitos e/ou produtores de conteúdo, e que por esse motivo necessitavam desenvolver uma infra-estrutura que apoiassem o comércio via rede envolvendo propriedade intelectual. Ao invés de propor um novo esquema de metadados, o <indecs> caracteriza-se por disponibilizar uma infra-estrutura comum que permite que transações entre esquemas relacionados a gêneros diferentes - como música, artigos de periódicos e livros - possam trocar informações, especialmente as concernentes à propriedade intelectual. Para tal, o <indecs> desenvolveu um núcleo comum de metadados obrigatórios (NISO, 2004b; Rust: Bide, 2000).

Várias organizações desenvolveram esquemas de metadados específicos sobre a infra-estrutura do <indecs>. Este é o caso do ONIX (Online Information Exchange) International Standard. Ele é um esquema de metadados baseado em XML,

desenvolvido por iniciativa da American Association of Publishers (AAP) em colaboração com a indústria livreira dos Estados Unidos e da Europa. Atualmente o ONIX é mantido e publicado pelo EDItEUR (http://www.editeur.org/). O ONIX for books é o padrão internacional para representar e comunicar informações sobre produtos da indústria livreira no formato eletrônico. Ele oferece elementos para registrar uma ampla gama de informações promocionais e de avaliação, bem como dados bibliográficos básicos e dados comerciais. ONIX for serials é uma família de formatos XML desenvolvido pela EDItEUR e a NISO para comunicar informações sobre periódicos e subscrições tendo como base muitos dos elementos definidos no ONIX for books (NISO, 2004b).

7.2.5 Objetos Visuais

Metadados usados para descrever objetos visuais, tais como pinturas ou esculturas, devem possuir requisitos específicos. Nessa direção a Art Information Task Force (AITF) desenvolveu uma infra-estrutura conceitual para descrever e acessar informações sobre obras de arte, arquitetura, materiais culturais, grupos e coleções de obras de arte e suas representações visuais denominado CDWA – Categories for Description of Works of Art (**http://www.getty.edu/research/conducting_research/standards/cdwa/**). O CDWA inclui 512 categorias e subcategorias, sendo que um subconjunto de categorias - considerada o núcleo - representa a informação mínima necessária para identificar sem ambigüidades e descrever um trabalho. Orientação, dimensões, condições, dedicatórias, tratamento de conservação e exposição são alguns dos elementos descritivos específicos e relevantes para o domínio das obras de arte (NISO, 2004b; Getty Foundation, 2006). CDWA Lite é um esquema XML para descrever registros do núcleo de categorias baseados no CDWA e no Cataloging Cutural Objects (CCO). Os registros do CDWA Lite são vocacionados para contribuição a catálogos coletivos e a outros repositórios por intermédio do Open Achives Initiatives (OAI) Harvesting Protocol (Getty Foundation, 2006).

Por outro lado, o VRA Core Categories – Visual Resources Association Data Standards Committee (http://www.vraweb.org/vracore3.htm) - está relacionado às coleções de registros visuais usados no ensino de história da arte e assuntos similares que não contêm obras originais, mas sim fotografias e slides dos originais. Isto significa que metadados para esses materiais têm que acomodar a descrição de níveis múltiplos de recursos relacionados, tais como a pintura original, um slide da pintura e uma imagem digitalizada do slide. Em resumo: o VRA Core Categories define um conjunto simples de metadados que pode ser usado para descrever uma obra, bem como as representações visuais dela.

7.2.6 Multimedia

Nos dias de hoje estão disponíveis em escala mundial uma grande quantidade de informação áudio-visual em formato digital que necessita de representação específica para serem adequadamente buscadas, identificadas e recuperadas. O ISO/IEC Moving Picture Expert Group (MPEG) vem desenvolvendo um conjunto de padrões para representação codificada de áudio e vídeo digital. Neste conjunto, dois padrões tratam de metadados, o MPEG-7 e o MPEG-21.

MPEG-7 – Interface de Descrição de Conteúdo Multimídia - define os elementos de metadados, a estrutura e os relacionamentos que são utilizados para descrever objetos

audiovisuais, incluindo figuras, gráficos, modelos em 3D, música, áudio, fala, vídeo e coleções multimídia. O MPEG-7 pode ser definido como uma interface para descrição de conteúdos multimídia que serve tanto para sistemas que atuam com usuários humanos, como para processamento automático. O padrão disponibiliza um conjunto de ferramentas de descrição (*Description Tools*), que incluem descritores (*Descriptors*) que definem a sintaxe e a semântica de cada elemento de metadados; e esquemas de descrição (*Description Schema*) que especificam a estrutura e a semântica dos relacionamentos entre os elementos. Com um olhar no futuro, espera-se que máquinas de busca possam utilizar os metadados de descrição MPEG-7 para identificar objetos áudio-visuais de maneira inteiramente nova, como por exemplo, tocar umas notas no teclado e recuperar uma lista de peças musicais que contém a seqüência de notas .(NISO, 2004b; Chella).

MPEG-21 foi desenvolvido para solucionar a necessidade de uma infra-estrutura que assegure a interoperabilidade de objetos digitais multimídias. Nesta direção, o padrão estabelece uma infra-estrutura normativa aberta para a disseminação e consumo desses objetos por todo os atores envolvidos. MPEG-21 é baseado em dois conceitos essenciais: a definição de uma unidade fundamental de distribuição e transação – Item Digital -, e o conceito de Usuários que interagem com os Itens Digitais. Dentro desse contexto, o MPEG-21 define a tecnologia necessária para dar apoio aos Usuários que desejam acessar, consumir, comercializar ou manipular Itens Digitais de forma eficiente, transparente e interoperável, e ainda considerando as permissões, direitos e propriedade intelectual (NISO, 2004b).

7.2.7 Estruturas de Metadados

O METS – Metadata Encoding and Transmission Standard (http://www.loc.gov/standards/mets/) - foi desenvolvido para preencher a necessidade de uma estrutura de dados padronizada para descrever objetos digitais complexos no contexto de uma biblioteca digital (NISO, 2004b). O METS é um esquema XML para a criação de instâncias de documento XML que expressa a estrutura de objetos digitais no âmbito de uma coleção, provê a codificação e o encapsulamento de metadados descritivos, administrativos e estruturais necessários para a recuperação, apresentação, gestão e preservação de objetos digitais, e ainda para o intercâmbio de informações entre repositórios. Adicionalmente, o METS oferece a possibilidade de associar um objeto digital com comportamentos ou serviços. Um documento METS consiste em sete principais seções: 1) Cabeçalho METS – contém metadados descrevendo o próprio documento METS; 2) Metadados descritivos – pode apontar para metadados descritivos externos ao documento METS (por exemplo: um registro MARC num OPAC ou um registro EAD num servidor web), ou conter metadados embutidos internamente ou ambos; 3) Metadados administrativos – oferece informações sobre como os arquivos foram criados e armazenados, direitos sobre propriedade intelectual, informações sobre o objeto original a partir do qual o objeto foi derivado, informações sobre a proveniência dos arquivos que compõe o objeto. Podem ser externos ou codificados internamente; 4) Seção de arquivos – relaciona todos os arquivos que compõem as versões eletrônicas do objeto digital; 5) Mapa estrutural – delineia uma estrutura hierárquica para o objeto da biblioteca digital que permite usuários navegar nele, vinculando os elementos dessa estrutura a arquivos de conteúdo e metadados referentes a cada elemento; 6) *Links* estruturais – permite aos criadores METS registrar a existência de *links* entre nós na hierarquia delineada no mapa estrutural. É de particular

valor quando o METS é utilizado para arquivar *sites*; 7) Comportamento – associa comportamentos executáveis com o conteúdo no objeto METS. O METS é mantido pela Library of Congress e desenvolvido como uma iniciativa da Digital Library Federation (Library of Congress, 2006; NISO, 2004b).

A RDF – Resource Description Framework (http://www.w3.org/RDF/)– , desenvolvida sob os auspícios da W3C - World Wide Web Consortium (http://www.w3.org/), estabelece uma infra-estrutura técnica que permite a codificação, intercâmbio e reuso de metadados estruturados. Trata-se de um modelo de dados para a descrição de recursos na Web que oferece mecanismos para integrar múltiplos esquemas de metadados. O modelo permite interoperabilidade de metadados entre aplicações que fazem intercâmbio de informações de forma automatizada na Web. Isto é realizado por meio da definição de mecanismos que suportem convenções consensuais de semântica, sintaxe e estrutura. Os metadados RDF podem ser usados em várias áreas de aplicação, por exemplo: na descoberta de recursos, possibilitando um melhor desempenho das máquinas de busca; na catalogação, descrevendo conteúdos; por agentes inteligentes na facilitação do compartilhamento e troca de conhecimento; além disso, constitui uma das bases da Web Semântica. A RDF usa XML como uma sintaxe comum para intercâmbio e processamento de metadados (Miller, 1998; NISO, 2004b).

8 PADRÕES DE INTEROPERABILIDADE

Interoperabilidade pode ser considerada como o processo contínuo de assegurar que sistemas, procedimentos e cultura de uma organização sejam gerenciados de modo a maximizar as oportunidades para intercâmbio e o reuso de informações. Considerando esta definição estabelecida por Paul Miller (Miller, 2000) e ratificada pela UKOLN (http://www.ukoln.ac.uk/), fica claro que a interoperabilidade está longe de depender somente de requisitos técnicos – como por exemplo, o uso de software e hardware compatíveis -, embora isso seja muito importante. Assegurar a plena interoperabilidade exige freqüentemente uma mudança radical na forma pela qual uma organização trabalha, se relaciona com as organizações parceiras, usuários e fornecedores e, especialmente, sua atitude diante dos problemas relacionados à informação. A interoperabilidade tem muitas faces, sendo que a mais visível é a interoperabilidade técnica , que tem como objeto o desenvolvimento contínuo de padrões e protocolos de comunicação, transporte, armazenamento e codificação de informações, tais como Z39.50, OAI-PMH, ISO-ILL, XML entre muitos outros, no entanto outros aspectos relevantes devem ser considerados, tais como:

a) Interoperabilidade semântica – está relacionado com a adoção de ferramentas comuns ou/e mapeáveis de representação da informação, como esquemas de metadados e tesauros;
b) Interoperabilidade política/humana – considera as implicações para a organização, equipe e usuários de tornar as informações mais amplamente disponíveis;
c) Interoperabilidade intercomunitária – enfoca a necessidade crescente de acesso a informações provenientes de um espectro amplo de fontes distribuídas por organizações e comunidades de natureza distintas. Geralmente exige o estabelecimento de fóruns para discussão e consenso em torno de práticas padronizadas;
d) Interoperabilidade legal – considera as exigências e as implicações legais de tornar livremente disponíveis itens de informação;

e) Interoperabilidade internacional – quando se atua em escala internacional é necessário contornar a diversidade de padrões e normas, os problemas de comunicação, as barreiras lingüísticas, as diferenças no estilo de comunicação e na falta de uma fundamentação comum.

8.1 Níveis de Interoperabilidade

Arms e seus colaboradores (Arms et al, 2002), no contexto do desenvolvimento da NSDL - National SMETE Digital Library -, identificam três níveis de interoperabilidade aplicáveis ao domínio das bibliotecas digitais: federação, *harvesting* (colheita automática de metadados) e *gathering* (agregação automática de informação). O nível mais alto, a federação, corresponde a mais potente forma de interoperabilidade, em contrapartida é a que exige maior ônus dos participantes Para se efetivar ela exige que um grupo de organizações concorde que seus serviços estejam em conformidade com um conjunto de especificações, geralmente selecionadas a partir de padrões formalizados. O principal desafio que se coloca na formação de federações é o esforço despendido por cada organização em implementar e manter atualizado todos os níveis dos acordos. As bibliotecas que compartilham registros de catálogos on-line usando o protocolo Z39.50, trabalham segundo o nível de federação. O ANSI/NISO39.50 (ISO 23950) é um protocolo de comunicação entre computadores que pode ser implementado sobre qualquer plataforma. Ele tem como propósito a pesquisa e a recuperação de informações. A implementação do protocolo permite que através de uma única interface seja possível o acesso uniforme a uma diversidade de fontes de informações heterogêneas de modo síncrono e quase transparente para o usuário-final (Healy, 2002).

Porém, as dificuldades de se criar grandes federações foi a principal motivação para busca de soluções menos onerosas para o estabelecimento de interoperabilidade entre bibliotecas digitais. Idéia subjacente é que os participantes concordem em despender um pequeno esforço que possibilite o compartilhamento de alguns serviços básico, sem que seja necessário o enquadramento a um conjunto completo de acordos. Nesta situação se enquadra o conceito de colheita automática de metadados (*metadata harvesting*), estabelecido pelo protocolo OAI-PMH – Open Archive Intiative – Protocol of Metadata Harvesting. Enquanto os serviços baseados em *havesting* são assíncronos e menos sofisticados do que os providos pelas federações, a sobrecarga sobre os participantes é consideravelmente menor. Como resultado, muito mais organizações estão optando por este tipo de interação, o que é provado pela rápida aceitação do OAI-PMH (Marcondes; Sayão, 2003b).

Mesmo que um determinado grupo de organizações não estabeleça nenhum grau formal de cooperação, um nível básico de interoperabilidade é ainda possível por meio de agregação automática de informações disponíveis publicamente, utilizando-se metabuscadores, robôs, máquinas de busca e ainda através dos protocolos que suportam *web services*. A agregação não requer essencialmente pouco ou nenhum esforço por parte dos participantes, entretanto oferece um grau baixo de interoperabilidade (Arms, 2002).

8.2 Interoperabilidade via *links* referenciais

Conforme foi analisado na seção 5, os *links* têm uma importância vital na interoperabilidade entre serviços oferecidos por bibliotecas digitais. Duas idéias inovadoras – uma expressa por um sistema e a outra por uma norma NISO - implementam o conceito de *links* referenciais e têm contribuído para surgimento de

serviços novos e surpreendentes. São eles o CrossRef (http://www.crossref.org). e o OpenURL Framework for Context-Sensitive Services (http://www.niso.org/standards/standard_detail.cfm?std_id=783) .

Todos reconhecem que as referências constituem um núcleo de vital importância nos periódicos científicos. É consenso que a citação num texto acadêmico, expresso por meio de referências bibliográficas, é um dos fundamentos básicos do sistema corrente de comunicação científica.. Com o surgimento dos periódicos eletrônicos, tornou-se possível o desenvolvimento de mecanismos que permitem o acesso imediato, via *links*, a trabalhos referenciados num artigo. Esta facilidade é hoje algo tão importante para os usuários que se transformou em um imperativo econômico para os editores científicos: ter hoje disponível nas publicações eletrônicas "*links* referenciais" tornou-se uma característica essencial. Reconhecendo a importância dos links referencias para os seus usuários e para os seus negócios, os editores científicos, numa iniciativa incomum de cooperação, estabeleceram as bases para um serviço colaborativo de *links* referenciais. Neste contexto que surgiu no ano 2000 o CrossRef. (Pentz, 2001; Brand, 2001). O CrossRef oferece uma infra-estrutura que sustenta um sistema de referências cruzadas via *links* referencias. Isto permite que um usuário clique numa referência citada em um periódico publicado pelo editor A, e assim seja conduzido diretamente ao conteúdo referenciado publicado num periódico do editor B. O Sistema se constitui essencialmente em uma base de dados – CrossRef Metadata Database (MDDB) - onde os editores associados depositam DOI´s e os URL´s correspondentes, além de um conjunto mínimo de metadados - título do periódico, volume, número, ISSN, paginação e autor principal - de artigos científicos publicados por eles. Esta submissão é feita em formato XML, segundo a *Document Type Definition* (DTD) disponível no *website* do CrossRef . Como parte do processo, os DOI´s e os URL´s dos artigos são registrados no Diretório Central DOI, operado pela DOI Foundation (http://www.doi.org) (Atkins, 2000).

Por outro lado, a norma OpenURL trata da resolução de links considerando o contexto do usuário. Quando um usuário clica sobre um *link* presente num recurso informacional, como por exemplo, uma citação num periódico eletrônico, ele é redirecionado sempre para a mesma versão do conteúdo – a versão *default* – geralmente publicada no *website* do editor. Isto acontece porque os *links* convencionais não levam em consideração a identidade do usuário, seu contexto institucional, suas preferências e direitos em termos de acesso e de serviços disponíveis para ele, conduzindo-os todos, indistintamente, para a mesma fonte. Se o servidor de *links* está informado sobre o contexto do usuário, ele será capaz de considerar a identidade do usuário quando da resolução dos metadados, direcionando-o para um serviço cujo acesso lhe seja o mais conveniente – a chamada "cópia apropriada" - e/ou apresentando um menu de opções de serviços. Isto configura uma solução aberta, sensível ao contexto. Dessa forma, a norma OpenURL define um protocolo para interoperabilidade entre um recurso de informação e um componente de serviço, denominado servidor de *links*, que oferece serviços de localização sensíveis ao contexto, através da interpretação de metadados bibliográficos, codificados segundo uma sintaxe própria, que descrevem uma publicação. Em resumo: A norma OpenURL especifica um formato padronizado para transportar metadados bibliográficos de objetos informacionais entre serviços de informação (Van de Sompel; Beit-Arie, 2001).

9 À GUISA DE CONCLUSÃO

O conceito de interoperabilidade está longe de ser uma novidade no domínio das bibliotecas. Desde sempre se soube que as bibliotecas não são ilhas e sempre precisaram, para cumprir bem o seu papel, trocar informações, estabelecer serviços cooperativos, intercambiar documentos. Toda uma estrutura global foi montada em torno da idéia do compartilhamento e da cooperação entre bibliotecas. Entretanto, com a consolidação e a concretização do conceito de bibliotecas digitais - que se localiza na interseção entre biblioteconomia, ciência da computação e tecnologias de rede -, aliados ao crescente interesse da indústria de conteúdos nas formas de disseminação dos repositórios digitais como meio de distribuição de seus produtos no ambiente de uma nova economia da informação, a interoperabilidade se torna um foco de grande interesse para muitos atores.

Vimos que a interoperabilidade depende fortemente de processos baseados em padrões abertos e bem documentados. Os projetos importantes de sistemas de bibliotecas digitais em todo o mundo – regionais, nacionais e internacionais -, aliados ao governo, empresários e profissionais da informação, têm se congregado em fóruns especiais para decidir sobre conjunto de padrões, protocolos, formatos e melhores práticas que possam ser adotadas em comum em seus projetos. Em nosso país, também e em nosso continente, essas ações têm sido pífias e de pouco alcance. Com a tecnologia disponível e barata, os repositórios digitais estão rapidamente se proliferando sem a perspectiva da integração e da interoperabilidade. É portanto urgente que se traduza a interoperabilidade humana e política em ações estruturantes e de regulamentação, para que finalmente possam ser criados novos repositórios digitais plenamente abertos e interoperáveis, e, não menos importante, para que se possam criar mecanismos para integrar também os já existentes.

REFERÊNCIAS BIBLIOGRÁFICAS

AIIM. *Frequently Asked Questions* (FAQs): ISO 19005-1: 2005 PDF/A-1. AIIM, July 2006. Disponível em <**http://www.aiim.org/documents/standards/19005-1_FAQ.pdf**>. Acessado em 05 maio 2007.

ALA COMITEE ON CATALOGUING: DESCRIPTION AND ACCESS. Task Force on Metadata. American Library Association, June 2000. Disponível em <**http://www.libraries.psu.edu/tas/jca/ccda/tf-meta6.html**>. Acessado em 04 maio 2007.

ALECRIM, Emerson. OpenDocument Format (ODF). Infowester, 2006. Disponível em <**http://www.infowester.com/odf.php**>. Acessado em 04 maio 2007.

ATKINS, Helen et al. Reference Linking with DOI´s. D-Lib Magazine, v.6, n.2, February 2000. Disponível em <**http://dlib.org/dlib/february00/02risher.html**>. Acessado em 04 maio 2007.

ARMS, William Y. *et al*. A Spectrum of Interoperability: the site for Science for Prototype for the NSDL. *D-Lib Magazine*, v.8, n.1, January 2002. Disponível em <http://www.dlib.org/dlib/january02/arms/01arms.html>. Acessado em 04 maio 2007.

ARMS, William Y. Key Concepts in the Architecture of the Digital Library. *D-Lib Magazine*, July 1995. Disponível em <http://www.dlib.org/dlib/July95/07arms.html>. Acessado em 05 maio 2007.

BERNERS-LEE, T., FIELDING, R., IRVINE, U.C., MASINTER, L. RFC2396 Uniform Resource Identifier (URI): Generic Syntax. Network Working Group, August 1988. Disponível em <http://rfc.net/rfc2396.html>. Acessado em 05 maio 2007.

BIGWOOD, David. Persistent links, one solution to a common problem. *HAL-PC Magazine*, June 1999. Disponível em <http://eprints.rclis.org/archive/00001991/01/links.html>. Acessado em 04 maio 2007.

BRAND, Amy. CrossRef Turns One. *D-Lib Magazine*, v.7, n.5, May 2001. Disponível em <http://www.dlib.org/dlib/may01/brand/05brand.html>. Acessado em 04 maio 2007.

BULLOCK, Alison. Preservation of digital information: issues and current status. Ottawa : National Library of Canada, April 22, 1999. Disponível em: <http://www.collectionscanada.ca/9/1/p1-259-e.html>. Acesso em 04 maio 2007.
CALIFORNIA DIGITAL LIBRARY. Digital Image Format Standards. California Digital Library, July 2001. Disponível em <http://chnm.gmu.edu/digitalhistory/links/pdf/chapter3/3.29b.pdf >. Acessado em 04 maio 2007.

CAPLAN, Priscilla, ARMS, William Y. Reference Linking for Journal Articles. *D-Lib Magazine*, v.5, n.4, April 1999. disponível em <http://www.dlib.org/dlib/july99/caplan/07caplan.html>. Acessado em 04 maio 2007.

CCSDS - CONSULTATIVE COMMITEE FOR SPACE DATA SYSTEM. Reference model for Open Archival Information System (OAIS): recommendation. Washington : CCSDS, 2002. 139p. Disponível em <http://public.ccsds.org/publications/archive/650x0b1.pdf>. Acesso em 04 maio 2007.

CHELLA, Marco Túlio. Sistema para Classificação e Recuperação de Conteúdo Multimídia Baseado no Padrão MPEG-7. Disponível em <http://www.nied.unicamp.br/~siros/doc/2232.pdf>. Acessado em 04 maio 2007.

CLEVELAND, Gary. Digital libraries: definitions, issues and challenges. IFLANET UDT Occasional Papers, March 1998. Disponível em <http://www.ifla.org/VI/5/op/udtop8/udtop8.htm>. Acessado em 04 maio 2007.

DACK, Diana. Persistent Identification Systems, part 1: background. National Library of Australia, May 2001. Disponível em <http://www.nla.gov.au/initiatives/persistence/PIpart1.html>. Acessado em 04 maio 2007.

DAY, Michael. Metadata for digital preservation: an update. *Ariadne*, v.22, Dec. 1999. Disponível em: <http://www.ariadne.ac.uk/issue22/metadata/>. Acesso em 04 maio 2007.

DIGITAL LIBRARY FEDERATION (2002). A Working Definition of Digital Library [1998]. Disponível em <http://www.diglib.org/about/dldefinition.htm>. Acessado em 04 maio 2007.

<EAD>. Development of the Encoded Archival description DTD. Library of Congress, 2006. Disponível em http://www.loc.gov/ead/eaddev.html>. Acessado em 04 maio 2007.

FLEISCHHAUER, Carl. Digital Formats for Content Reproduction. Washington, DC: Library of Congress, 1998. Disponível em <http://lcweb2.loc.gov/ammem/formats.html>. Acesso em 04 maio 2007.

GETTY FOUNDATION. Categories for the Description of Works of Art. Getty Foundation, 2006. Disponível em <http://www.getty.edu/research/conducting_research/standards/cdwa/introduction.html >. Acessado em 04 maio 2007.

GÓMEZ, Alejandro Delgado. Introducción a Encoded Archival Description (EAD): Mitos y oportunidades. Disponível em <http://sapp.telepac.pt/apbad/congresso8/convidado1.pdf >. Acessado em 12 dez. 2006.

HAIGH, Susan. A glossary of digital library: standards, protocols and format. Libraries and Archives Canada, 1998. Disponível em <http://www.collectionscanada.ca/9/1/p1-253-e.html>. Acessado em 04 maio 2007.

HEALY, Leigh Watson. Z39.50 – A primer on the protocol. Bathesda, MD : NISO Press, May 2002. Disponível em <http://www.niso.org/standards/resources/Z3950_primer.pdf>. Acessado em 04 maio 2007.

HEDSTROM, Margaret. Digital preservation: a time bomb for digital libraries. *Computer and the Humanities*, v.31, n.3, p.189-202, 1997/1998. Disponível em: <http://www.uky.edu/~kiernan/DL/hedstrom.html>. Acesso em 04 maio 2007.

IEEE. WG12: Learning Object Metadata. Disponível em http://ltsc.ieee.org/wg12/. Acessado em 04 maio 2007.

ISO - INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. Overview of the ISO System. ISO, 2006. Disponível em <http://www.iso.org/iso/en/aboutiso/introduction/index.html#one >. Acessado em 04 maio 2007.

KOEHLER, W. Web page change and persistence: a four-year longitudinal study. *Journal of the American Society for Information Science and Technology*, v.53, n.2, p.162-171, 2002. (DOI:10.1002/asi.10018).

LAVOIE, Brian. Meeting the challenges of digital preservation: the OAIS reference model. OCLC –Newsletter, n.243, p.26-30, Jan./Feb. 2000. Disponível em: <http://www.oclc.org/research/publications/archive/2000/lavoie/>. Acesso em 04 maio 2007.

LEE, Kyong-Ho *et al*. The State of the Art and Practice in Digital Preservation. *Journal of Research of the National Institute of Standards and Technology*, v.107, n.1, p. 93-106, January-February 2002. Disponível em: <http://nvl.nist.gov/pub/nistpubs/jres/107/1/j71lee.pdf.>. Acesso em 04 maio 2007.

LIBRARY OF CONGRESS. MARC XML Design Considerations. Library of Congress, 2004. Disponível em <http://www.loc.gov/standards/marcxml/marcxml-design.html>. Acessado em 04 maio 2007.

LIBRARY OF CONGRESS. METS: Introdução & Tutorial. Library of Congress, 2006. Disponível em < http://www.loc.gov/standards/mets/METSOverview.v2_port.html>. Acessado em 04 maio 2007.

MARCONDES, Carlos Henrique, SAYÃO, Luis Fernando. The SciELO Brazilian Scientific Journal Gateway and Open Archives: A Report on the Development of the SciELO-Open Archives Data Provider Server. *D-Lib Magazine*, v.9, n.3, March 2003a. Disponível em: <http://www.dlib.org/ dlib/march03/marcondes/03marcondes.html>. Acesso em: 04 maio 2007.

MARCONDES, Carlos Henrique, SAYÃO, Luis Fernando. Brazilian Digital Library of Theses and Dissertations. *The International Information & Library Review*, n.35, v.2-4, June-December 2003b, p.265-279.

MILLER, Paul. Interoperability: what is it and why should I want it. Ariadne, n.24, June 2000. Disponível em <http://www.ariadne.ac.uk/issue24/interoperability/>. Acessado em: 04 maio 2007.

MILLER, Eric. An Introduction to the Resource Description Framework. D-Lib Magazine, May 1998. Disponível em <http://www.dlib.org/dlib/may98/miller/05miller.html>. Acessado em 04 maio 2007.

MOATS, R. RFC 2141 URN Syntax. Network Working Group, May 1997. Disponível em <http://www.rfc-archive.org/getrfc.php?rfc=2141>. Acessado em 04 maio 2007.

MPEG ACHIEVEMENTS. Disponível em <http://www.chiariglione.org/mpeg/achievements.htm>. Acessado em 26/01/2007.

NISO. Framework Advisory. A framework of Guidance for Building Good Digital Collection. Bathesda, MD : National Information Standards Organization, 2004a. Disponível em <http://www.niso.org/framework/framework2.pdf >. Acessado em 04 maio 2007.

NISO. Understanding Metadata. Bethesda, MD : NISO Press, 2004b. Disponível em <http://www.niso.org/standards/resources/ UnderstandingMetadata.pdf>. Acessado em 04 maio 2007.

NOERR, Peter. The Digital Library Toolkit. Sun Microsystems. 3rd edition. Santa Clara, CA.: Sun Microsystems, January 2003. Disponível em <http://www.sun.com/products-n-solutions/edu/whitepapers/pdf/digital_library_toolkit.pdf >. Acessado em 04 maio 2007.

OGBUJI, Uche. Thinking XML: Learning Objects Metadata. IBM, 2003. Disponível em < http://www-128.ibm.com/developerworks/xml/library/x-think21.html>. Acessado em 04 maio 2007.

PAYETTE, Sandra. Persistent identifiers on the digital terrain. *RLG Diginews*, v.2, n.2, April 15, 1998. Disponível em <http://www.rlg.org/preserv/diginews/diginews22.html#Identifiers>. Acessado em 04 maio 2007.

PDF TOOLS AG. PDF/A The Basics. PDF Tools AG, 2006. Disponível em <http://www.pdf-tools.com/public/downloads/whitepapers/whitepaper-pdfa.pdf>. Acessado em 05/01/2007.

PENTZ, Ed. A Collaborative Linking Network. *Issues in Science and Technology Librarianship*, n.29, Winter 2001. Disponível em <http://www.library.ucsb.edu/istl/01-winter/article1.html>. Acessado em 04 maio 2007

PITTI, Daniel V. Encoded Archival Description: an introduction and overview. *D-Lib Magazine*, v. 5, n.11, November 1999. Disponível em <http://www.dlib.org/dlib/november99/11pitti.html>. Acessado em 04 maio 2007.

PRINCETON UNIVERSITY LIBRARY. Imaging Standards: about. Disponível em <http://diglib.princeton.edu/?_xq=html&_xsl=imaging.xsl>. Acessado em 04 maio 2007.

RUST, Godfrey; BIDE, Mark. The <indecs> Metadata Framework: Principles, Models and Data Dictionary. <indecs>, June 2000. Disponível em <http://www.indecs.org/pdf/schema.pdf >. Acessado em 05/01/2006.

SAYÃO, Luis Fernando. Preservação digital no contexto das bibliotecas digitais. In: Marcondes , C. H; Kuramoto, H.; Toutain, Lidia Brandão; Sayão, Luis Fernando. (Org.). Bibliotecas digitais: saberes e práticas. Salvador/Brasília: UFBA/IBICT, 2006, p. 115-149.

SHAFER, Keith; WEIBEL, Stuart; JUL, Erik; FAUSEY, Jon. Introduction to persistent Uniform Resource Locators. OCLC Online Computer Library Center, 1996. Disponível em <http://www.isoc.org/inet96/proceedings/a4/a4_1.htm>. Acessado em 04 maio 2007.

SHIRI, Ali. Digital library research: current developments and trends. *Library Review*, v. 32, n.5, p.198-202, 2003. Disponível em <http://eprints.rclis.org/archive/00001521/02/ASLRcolumn.pdf >. Acessado em 04 maio 2007.

SOLLINS, K; MASINTER, L. RFC 1737 Functional Requirements for Uniform Resource Names. Network Working Group, December 1994. Disponível em <http://www.ietf.org/rfc/rfc1737.txt>. Acessado em 04 maio 2007.

TASK FORCE ON THE ARCHIVING DIGITAL INFORMATION. Preservation digital information: report of the Task Force on Archiving Digital Information. Washington, DC. : Commission on Preservation and Access, 1996. Disponível em: <http://www.rlg.org/ArchTF>. Acesso em 04 maio 2007.

UNIVERSITY OF VIRGINIA LIBRARY. TEI Guideline for Electronic Text Encoding and Interchange. 2003. Disponível em <http://etext.lib.virginia.edu/standards/tei/teip4/teip4-tocs1.html>. Acessado em 04 maio 2007.

UNICODE FOUNDATION. O que é o Unicode?. Unicode Foundation, 2005. Disponível em <http://www.unicode.org/standard/translations/portuguese.html>. Acessado em 04 maio 2007.

VAN de SOMPEL, Hebert, BEIT-ARIE, Oren. Open linking in the scholarly information environment using the Open URL framework. *D-Lib Magazine*, v.7, n.3, March 2001. Disponível em <http://www.dlib.org/dlib/march01/vandesompel/03vandesompel.html>. Acessado em 04 maio 2007.

WEIBEL, Stuart, JUL, Eric, SHAFER, Keith. PURLs: Persistent Uniform Resource Locator. OCLC Online Computer Library Center, 1995. Disponível em <http://purl.oclc.org/docs/new_purl_summary.html>. Acessado em 04 maio 2007.

ABSTRACT

Interoperability is one of the main issues in creating a networked system of digital libraries. However, the interoperability as the way to accomplish data exchange and service collaboration requires adoption of a set of open standards covering all digital repository processes. The aim of this document is to revise the most important standards, protocols and the best pratices that form the framework to an open and fully interoperable digital library.

KEYWORDS: Digital Libraries, Interoperability. Standards. Protocols.

Originais recebidos em: 15/03/2007
Texto aprovado em 15/06/2007