SInBAD – Sistema Integrado de Biblioteca e Arquivo Digital

Pedro Almeida[1], Marco Fernandes[1], Miguel Alho[1], Joaquim Arnaldo Martins[2], Joaquim Sousa Pinto[2]

[1] IEETA, Universidade de Aveiro
3810-193 Aveiro, Portugal
{pma, marcopsf, alho}@ieeta.pt

[2] DET, Universidade de Aveiro
3810-193 Aveiro, Portugal
{jam, jsp}@det.ua.pt

Resumo. O SInBAD – Sistema Integrado de Biblioteca e Arquivo Digital, é um projecto que está a ser desenvolvido na Universidade de Aveiro. O objectivo deste projecto é construir um sistema de informação capaz de armazenar os diferentes tipos de documentos que são propriedade ou são produzidos na Universidade de Aveiro, como por exemplo livros, teses, dissertações, fotografias, vídeos, músicas, etc. Estes conteúdos são acedidos através de diferentes interfaces de acordo com o tipo de informação que está a ser visualizada. Para além da questão crucial do arquivo e armazenamento, estabeleceu-se também como objectivo a criação de um portal de pesquisa para todos estes conteúdos – através de um único ponto de pesquisa, o utilizador pode obter informação sobre livros, teses, vídeos, músicas, etc., que digam respeito a um determinado assunto. A arquitectura do SInBAD é composta por vários subsistemas. Um aspecto a realçar, é que a o componente de acesso à informação foi desenhada de forma a permitir ter um armazenamento de informação distribuído. Ao nível de comunicação entre os diferentes componentes do sistema utilizam-se tecnologias baseadas em XML, nomeadamente WebServices. O modelo de descrição adoptado para os vários tipos de informação tem por base o conjunto de elementos básicos definido pelo grupo Dublin Core. Acontece que esta descrição pode não ser suficiente para caracterizar devidamente os registos, daí utilizar-se para cada descrição, uma extensão com um standard mais apropriado para o material descrito. As descrições são armazenadas no formato XML, tendo em vista a interoperabilidade e a extensibilidade que esta linguagem suporta por natureza.

PALAVRAS-CHAVE

Bibliotecas Digitais, Arquivos Digitais, Sistemas de Informação.

1. INTRODUÇÃO

O SInBAD, Sistema Integrado de Biblioteca e Arquivo Digital, é um projecto que está a ser desenvolvido para a Universidade de Aveiro, financiado por o programa Aveiro-Digital 2003-2006 [1]. Este projecto tem como finalidade construir um sistema de informação que armazene e disponibilize livros, teses, fotografias, vídeos, músicas, etc., no formato digital, sendo que a informação armazenada no sistema diz respeito à vida académica da Universidade de Aveiro. Este repositório de informação vai servir para armazenar, por exemplo, as fotografias de início do ano lectivo, discursos proferidos por personalidades importantes da Universidade de Aveiro, as teses e dissertações apresentadas na Universidade de Aveiro, etc. Para além do registo em formato digital é também necessário armazenar a descrição do mesmo, onde deve constar o nome do autor, assunto, editor, etc. Mas como nem todos os registos são do mesmo formato, é necessário classificar cada documento de acordo com o tipo de registo que está a ser armazenado. Por exemplo, no caso de uma fotografia, convém fazer a distinção entre imagens a cores ou escala cinza, mas no caso de uma música esta descrição não faz sentido. Este é somente um dos aspectos que é necessário ter em consideração, entre vários que vamos apresentar neste artigo.

Uma das inovações associadas a este sistema é a capacidade de armazenar qualquer tipo de registo e disponibilizar mecanismos de integração de conteúdos. Isto porque a informação que consta no SInBAD está dividida, até agora, em três grandes grupos: Biblioteca, Arquivo e Jazz. Estes três grandes grupos reflectem um pouco a organização interna da instituição ao nível da preservação e disponibilização de conteúdos, pois a Biblioteca da Universidade de Aveiro é que está responsável pela organização e classificação dos livros e teses, entre outra grande variedade de documentos, o Arquivo da Universidade de Aveiro está responsável por organizar as fotografias e vídeos referentes à vida académica e, finalmente, o Jazz representa uma iniciativa que está a ser desenvolvida na Universidade, iniciativa esta que pretende divulgar um vasto acervo de Jazz pertencente à Universidade de Aveiro. A integração destes três sistemas é feita através do SInBAD, um portal, um único ponto de pesquisa, onde um utilizador pode fazer pesquisas nos conteúdos da Biblioteca, Arquivo e Jazz. Ao fazer uma pesquisa no SInBAD o utilizador vai obter uma listagem de teses, livros, fotografias, vídeos, músicas, etc., que apresentem conteúdos relacionados com a pesquisa efectuada. Cada registo é apresentado através de uma interface própria, pois cada um destes catálogos tem um subsistema que apresenta a informação num formato amigável. Este artigo apresenta também os objectivos do projecto SInBAD, a arquitectura do sistema e algumas conclusões sobre a investigação desenvolvida até ao momento.

2. TRABALHO RELACIONADO

No artigo [2], Robert Tansley et. al. apresentam o DSpace [3]: um sistema *open-source* que actua como um repositório para material digital educacional e de investigação produzido por uma organização ou instituição. Este sistema foi desenvolvido numa parceria entre a HP e a biblioteca do MIT e foi construído tendo

em conta todas as funcionalidades necessárias ao bom funcionamento de um serviço de repositório digital. Este serviço deve ser um serviço "vivo", sendo o alicerce para as funcionalidades de um repositório e particularmente para resolver as preocupações de armazenamento a longo prazo. Todos os registos que estão armazenados no DSpace podem ser livremente acessíveis, não existindo restrições no acesso à informação. Os aspectos funcionais do DSpace podem ser resumidos da seguinte forma:

1. É definido um modelo de dados para organizar a informação de uma forma básica.
2. Metadados de vários tipos são armazenados pelo sistema.
3. O sistema armazena informação acerca dos utilizadores do sistema. Alguns utilizadores podem não ser humanos mas sim outros sistemas computacionais.
4. Enquanto muito do esforço visa facilitar o acesso ao material digital de uma instituição, simplesmente permitir acesso livre total nem sempre é aceitável. Funções adicionais como depositar ou avaliar devem ser restringidas a determinados indivíduos. Daí que o sistema inclua uma funcionalidade de autorização.
5. O sistema tem que ser capaz de aceitar material submetido, um processo definido com ingestão.
6. Algumas comunidades podem requerer que o material ou metadados relacionados que são submetidos para arquivo sejam verificados por indivíduos designados. Este processo chama-se *workflow*.
7. Tendo em vista que o material armazenado pode ser citado e acedido usando informação de uma citação, um sistema de gestão de identificadores é usado para atribuir identificadores únicos e persistentes aos objectos arquivados.

A nível de funcionalidades de pesquisa e descoberta o DSpace permite que os utilizadores acedam ao conteúdo através de uma referência externa, por exemplo um identificador, procurando por uma ou mais palavras chave, navegando pelos índices de títulos, data e autor. Convém também salientar que o DSpace implementa o protocolo OAI-PMH [5], expondo os metadados Dublin Core dos registos que são públicos.

A arquitectura do DSpace é baseada num modelo de três camadas. Utiliza uma base de dados relacional para armazenar a informação sob a forma de uma cadeia de bits, mantendo os metadados acessíveis pelo sistema. As funcionalidades de indexação, autorização, gestão de utilizadores, etc., estão localizadas na camada lógica, e esta disponibiliza um conjunto de API's para comunicar com aplicações externas. Outra característica importante é que utiliza um serviço do ORNI Handle Server [6] para atribuir identificadores únicos ao registos que armazena.

3. OBJECTIVOS

Durante a fase de análise do projecto analisou-se a possibilidade de adaptar o DSpace para construir o SInBAD, mas devido a algumas limitações do DSpace optou-se por desenvolver um sistema de raiz. Comparativamente ao DSpace, o SInBAD suporta as funcionalidades 2,3,4,5 e 6 apresentadas na secção 2 e definimos também os seguintes requisitos:

1. Utilizar XML [8] e tecnologias baseadas em XML para armazenar e disponibilizar o acesso à informação existente no sistema.
2. Adoptar uma arquitectura distribuída para suportar diversos repositórios em vez de um repositório único.
3. Suporte para descrições multinível.
4. Estrutura de organização de informação especificada em cada subsistema.
5. Apresentação dos documentos de uma forma estruturada.

4. ARQUITECTURA

Os objectivos mencionados, foram também definidos tendo em consideração a integração do SInBAD com os vários sistemas que existem na Universidade de Aveiro, nomeadamente o sistema Aleph [9] da Biblioteca, onde são armazenadas as descrições dos livros, revistas, etc., e o sistema PACO [10] onde são armazenadas as informações sobre os alunos, como por exemplo as disciplinas a que estão inscritos.

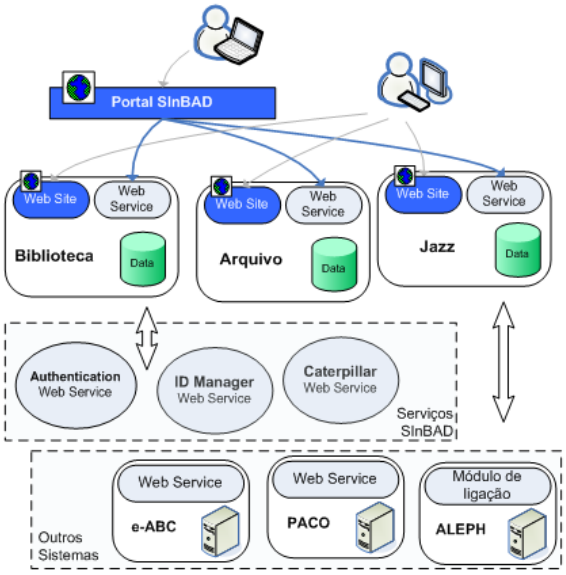

Figura 1 - Arquitectura do SInBAD

A arquitectura do SInBAD utiliza WebServices [11] por forma a permitir uma fácil integração dos subsistemas do SInBAD em outros portais de informação. A Figura 1 ilustra a arquitectura do sistema.
O SInBAD em si é composto por vários subsistemas, nomeadamente o subsistema da Biblioteca, o subsistema do Arquivo e o subsistema do Jazz.

4.1 Subsistemas

Cada um dos subsistemas do SInBAD poderia existir isolado, daí que para cada subsistema exista um Web Site individual. O bloco que aparece com o portal SInBAD faz a integração de todos os subsistemas existentes. O portal funciona como um ponto de acesso a partir do qual se podem fazer pesquisas em todos os subsistemas que existem no portal, Biblioteca e Arquivo e Jazz. Desta forma, caso sejam criados mais subsistemas dentro do SInBAD, como por exemplo um subsistema de informação para museus, desde de que este seja construído segundo a arquitectura do SInBAD facilmente se integraria este novo subsistema no portal, permitindo fazer pesquisas simultâneas na Biblioteca, Arquivo, Jazz e Museu Digital.

Figura 2 - Interface do SInBAD

Após efectuar uma pesquisa no portal SInBAD este devolve uma listagem com documentos dos vários subsistemas, mas quando o utilizador selecciona um desses documentos é a interface do respectivo subsistema que vai mostrar a informação. Desta forma garante-se que a informação é apresentada de uma forma estruturada ao

utilizador, isto é, da melhor forma possível, uma vez que cada subsistema pode implementar as suas regras de apresentação dos conteúdos.
Já foi desenvolvido um protótipo do SInBAD e a Figura 2 ilustra o resultado de uma pesquisa efectuada no sistema. Como se pode observar, o sistema devolve registos de teses, pertencendo estas ao subsistema da Biblioteca Digital, e registos do Arquivo Fotográfico. Este protótipo permite ver perfeitamente o tipo de resultados que se pode obter com este sistema e a integração de pesquisas em vários sistemas de informação.

4.2 Arquitectura dos subsistemas

Cada subsistema é composto por uma série de componentes que formam a arquitectura dos subsistemas, como se pode observar na Figura 3. Esta divisão por subsistemas permite que cada um defina as regras de consulta e as regras de acesso aos seus conteúdos. O bloco *query manager* é responsável por implementar as funções de acesso aos dados, nomeadamente consultar, pesquisar, eliminar e alterar registos.
Um aspecto a destacar é que cada subsistema implementa o seu próprio modelo de organização e descrição de dados. Este factor é importante, pois os arquivistas, bibliotecários, museólogos, etc., não trabalham da mesma forma e cada um usa os seus próprios standards de descrição e regras de organização da informação. Isto permite que os subsistemas se adaptem aos requisitos do utilizador e não que o utilizador se adapte ao modo de funcionamento dos sistemas.
Cada subsistema contém um módulo de gestão de direitos, o bloco *Rights Manager*. Existem diversos perfis de utilizadores do SInBAD e para cada perfil é necessário verificar os direitos do utilizador. Por exemplo se um utilizador tentar aceder a uma tese em formato digital tem que estar autenticado como aluno, docente ou funcionário da Universidade de Aveiro, caso contrário não pode ver o documento em formato digital. Cada documento inserido no SinBAD tem associada uma descrição dos direitos de visualização por perfil de utilizador.

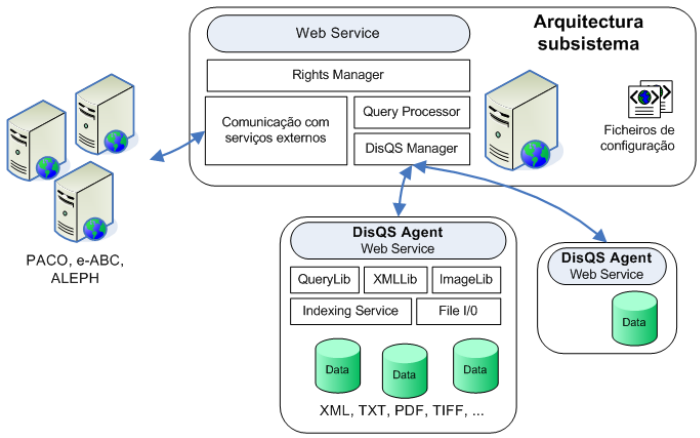

Figura 3 - Arquitectura dos subsistemas

4.2.1 Acesso e armazenamento de informação

As funções de acesso e manipulação de dados são geridas através de Web Services, aumentando desta forma a interoperabilidade do sistema. Cada subsistema deve implementar o seu próprio Web Service. Sempre que um novo documento é inserido num dos subsistemas o módulo *ID Manager Web Service* atribui-lhe um identificador único no sistema.
O armazenamento da informação, isto é, o documento propriamente dito e os metadados associados ao documento, é efectuado pelo módulo *DisQS Manager*, um sistema de armazenamento distribuído. O *DisQS Manager* faz a gestão de um conjunto agentes que armazenam as descrições em formato XML e os documentos, sendo que cada documento pode ser armazenado em um ou mais agentes DisQS, sempre com o mesmo identificador que foi atribuído pelo módulo *Id Manager*. Estes agentes, funcionam como um conjunto de diferentes repositórios que actuam como um repositório virtual único. Este repositório virtual único, tem como objectivo disponibilizar um conjunto de funcionalidades que visam aumentar a fiabilidade e desempenho do sistema global, conforme se descreve detalhadamente em [12]. Neste módulo a informação é armazenada de forma redundante, existindo um mesmo ficheiro em diferentes repositórios. Quando é efectuada uma consulta ao *DisQS Manager*, este agrupa os resultados dos vários agentes e elimina os itens repetidos.

4.2.2 Transformação de dados

O SInBAD suporta o armazenamento de diversos tipos de ficheiros, mas os formatos preferenciais são o PDF [13] e o TIFF [14]. Ambos estes formatos têm um problema, é que não são suportados directamente por os browsers web. Sendo assim, muitas vezes é necessário converter imagens para um formato diferente daí a existência do *Caterpillar Web Service*, um módulo que tem um conjunto de funcionalidades associadas, como por exemplo a conversão de imagens TIFF para PNG, e a extracção de texto e imagens de documentos PDF. Depois de transformadas as imagens são colocadas numa cache, por forma a que em novos pedidos não seja necessário converter novamente a imagem, aumentando desta forma o desempenho do sistema.

4.2.3 Interfaces com outros sistemas

A arquitectura dos subsistemas também contempla interfaces de comunicação com outros sistemas da Universidade, nomeadamente o PACO e o ALEPH, por forma a reutilizar informação da catalogação das teses, livros, etc., e comunicar com o sistema da secretaria da Universidade por forma a saber quais as disciplinas em que os alunos estão inscritos, quais as disciplinas de que um docente é responsável, o plano de estudos da disciplina, etc., sendo depois esta informação utilizada para atribuir as permissões de acesso à informação e para criar serviços de valor acrescentado. Por exemplo, após fazer *login* no SInBAD, um aluno pode saber quais as disciplinas a que está inscrito e pode aceder directamente à bibliografia da disciplina em formato digital.

5. MODELO DE DESCRIÇÃO DOS REGISTOS

Cada subsistema do SInBAD utiliza diferentes standards de descrição dos documentos de acordo com as características do material armazenado. No entanto, para além da descrição pormenorizada existe também uma descrição elementar para cada um dos registos segundo a norma Dublin Core. O formato de descrição Dublin Core é composto por um conjunto reduzido de elementos que existem em todos os tipos de registos: o título, autor, assunto, descrição, editor, etc.. O modelo de descrição adoptado no caso da Biblioteca Digital utiliza Dublin Core para fazer a descrição base e DC-Lib [15] para fazer a descrição pormenorizada. No caso do Arquivo Digital existem três tipos de registos: textos, imagens e audiovisuais. Cada um destes registos deve utilizar standards de descrição diferentes de acordo com a natureza do documento. Caso se trate de um registo audiovisual a descrição é armazenada segundo a norma MPEG-7 [16], no caso de ser uma fotografia a descrição é armazenada segundo a norma VRACore [17] e no caso de se tratar de um documento de arquivo usa-se a norma ISAD(G) [18]. Estas descrições armazenam as características específicas dos registos, sendo que o resto da informação é armazenada segundo a norma Dublin Core. A Figura 4 ilustra o modelo de descrição adoptado para o Arquivo Digital: uma descrição base segundo a norma Dublin Core e a extensão dos restantes standards de acordo com a natureza do documento associado.

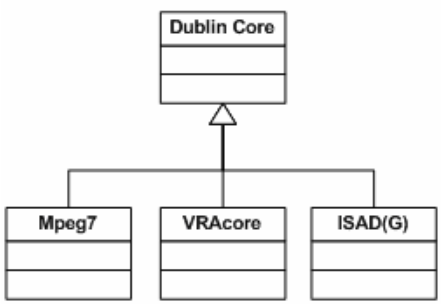

Figura 4 - Modelo de descrição do Arquivo Digital

Uma vez que todos os registos têm uma descrição básica Dublin Core, independentemente de pertencerem ao Arquivo ou à Biblioteca Digita e mesmo independentemente do formato, pesquisando por um dos campos Dublin Core pode-se obter uma listagem de registos pertencentes a qualquer um dos sistemas, sendo que este registo pode ser um vídeo, um livro, uma tese, etc.. O *DisQS Manager* é que indica de onde veio o registo e que tipo de informação é que contém, isto é, se é um vídeo, uma fotografia ou um documento.
A Figura 5 representa um exemplo de uma descrição de um cartaz de Jazz. Como se pode observar, existe um conjunto de elementos Dublin Core que descrevem as características gerais do cartaz e alguns elementos VRA Core que descrevem características específicas, como por exemplo as dimensões do cartaz.

```
<document xmlns:sinbad="http://sinbad.ua.pt" xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:dclib="http://purl.org/dc/terms/" xmlns:dcterms="http://purl.org/dc/terms2/"
xmlns:vracore="http://www.vraweb.org/vracore3.htm"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.ua.pt/SInBAD/Cartazes.xsd"
xmlns="http://www.ua.pt/SInBAD/Cartazes">
      <dc:date>2002</dc:date>
      <dc:description>Texto do cartaz: 25 Out. a 2 Nov., Grande Auditório do Centro de Arte e
    Espectáculos. Identificação dos músicos/intérpretes dos diversos concertos a realizar e
    respectivas datas.</dc:description>
      <vracore:Measurements.Dimensions>30 x 21 cm.</vracore:Measurements.Dimensions>
      <vracore:IdNumber.FormerRepository>00190216</vracore:IdNumber.FormerRepository>
      <dc:language>por</dc:language>
      <dc:publisher>Câmara Municipal</dc:publisher>
      <vracore:subject>Cartazes de concertos</vracore:subject>
      <dc:title>FozJazz 2002 : 1º Festival Internacional de Jazz da Figueira da Foz</dc:title>
      <dcterms:spatial>Figueira da Foz</dcterms:spatial>
      <dc:identifier
    xsi:type="dcterms:URI">http://sinbad.ua.pt/cartazes/CEJ_JD_CT_I_18</dc:identifier>
      <dc:subject>Música popular</dc:subject>
      <dc:subject>Jazz</dc:subject>
      <dc:subject>Concertos de jazz</dc:subject>
      <dc:type>Cartaz</dc:type>
      <dc:rights><sinbad:visible/></dc:rights>
      <sinbad:changed date="21-12-2005 15:23:42">jazzedit</sinbad:changed>
    </document>
```

Figura 5 – Exemplo de uma descrição XML de um cartaz

6. CONCLUSÃO

Da análise do SInBAD no seu estado actual podemos concluir que o objectivo principal, integrar o sistema de Biblioteca e Arquivo Digital da Universidade de Aveiro foi alcançado. Como vantagens deste sistema sobre os demais existentes destacamos os factos de usar standards baseados em XML, uma linguagem extensível, para o armazenamento das descrições, Web Services para a comunicação entre sistemas por forma a garantir a interoperabilidade com outros sistemas existentes na Universidade, a utilização de subsistemas por forma a aumentar a flexibilidade no que diz respeito à apresentação e organização da informação.

Utilizando a arquitectura SInBAD é possível criar um grande repositório com toda a informação de uma instituição, integrando diversas fontes de informação num único portal. Neste caso concreto foram integrados a Biblioteca, o Arquivo Digital e o Jazz, mas facilmente poderiam ser agregados mais subsistemas.

O modelo de descrição adoptado para o sistema permite que os subsistemas utilizem os standards de descrição adequados. Independentemente da natureza dos documentos utiliza-se sempre uma descrição Dublin Core que permite fazer pesquisas em todos os registos do SInBAD, obtendo deste modo uma funcionalidade de

integração básica, mas com valor acrescentado para grandes repositórios de informação heterogénea.

REFERÊNCIAS

[1] Aveiro Digital, 2004. http://www.aveiro-digital.pt, último acesso em Novembro 2004.

[2] Tansley, R. et al, 2003, The DSpace Institutional Digital Repository System: Current Functionality, *Proceedings of the Joint Conference on Digital Libraries (JCDL'03)*, Houston, USA.

[3] MIT, 2003. *DSpace Federation*, http://www.dspace.org/, último acesso em Março de 2005.

[4] Dublin Core Metadata Initiative, 2004. *Dublin Core Metadata Element Set, Version 1.1: Reference Description*, http://www.dublincore.org/documents/dces/, último acesso em Novembro de 2005.

[5] OAI, 2005. *The Open Archives Initiative Protocol for Metadata Harvesting*, http://www.openarchives.org/OAI/openarchivesprotocol.html, último acesso em Novembro de 2005.

[6] S. Sun, L. Lannom, B. Boesch CNRI, 2003. *Handle System Overview*, http://www.ietf.org/rfc/rfc3650.txt, último acesso em Novembro de 2005.

[7] University of Waikato, 2005. *The Greenstone Digital Library Software*, http://www.greenstone.org, último acesso em Março de 2005.

[8] W3C, 2005. *Extensible Markup Language (XML)*, http://www.w3.org/XML/, último acesso em Novembro de 2005.

[9] Universidade de Aveiro, 2005. *Universidade de Aveiro - Serviços de Documentação*, http://www.doc.ua.pt/, último acesso em Novembro de 2005.

[10] Universidade de Aveiro, 2005. *Portal Académico Online*, http://paco.ua.pt/, último acesso em Novembro de 2005.

[11] W3C, 2004. *Web Services Architecture*, http://www.w3.org/TR/ws-arch/, último acesso em Novembro de 2005.

[12] Almeida, P. et al, 2005, DisQS - Distributed Query System Based on Web Services for Digital Librarie, *Proceedings of the First International Conference on Internet Technologies and Applications (ITA 05)*, Wrexham, UK.

[13] Adobe Developers Association, 2005. *PDF Reference*, http://partners.adobe.com/public/developer/pdf/index_reference.html, último acesso em Novembro de 2005.

[14] Adobe Developers Association, 1992. *TIFF Revision 6.0 Final*, http://partners.adobe.com/public/developer/en/tiff/TIFF6.pdf, último acesso em Novembro de 2005.

[15] Dublin Core Metadata Initiative, 2004. *DC-Library Application Profile (DC-Lib)*, http://www.dublincore.org/documents/library-application-profile/, último acesso em Novembro de 2005.

[16] International Organization for Standardization, 2003, *MPEG-7 Overview (version 9)*, http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm, último acesso em Novembro de 2005.

[17] Visual Resources Association Data Standards Committee, 2002. *VRA Core Categories, Version 3.0*, http://www.vraweb.org/vracore3.htm, último acesso em Novembro de 2005.

[18] ICA - International Council on Archives, 1999, *ISAD(G): General International Standard Archival Description*, Second Edition, Stockholm, Sweden, http://www.ica.org/biblio/cds/isad_g_2e.pdf, último acesso em Novembro de 2005.