BEATRIZ DO CARMO LANGIANO

UM MECANISMO PARA AUTOMATIZAR A CRIAÇÃO DOS METADADOS DAS IMAGENS DE BIBLIOTECAS DIGITAIS E PROVER BUSCAS POR CONTEÚDO

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre. Programa de Pós-Graduação em Informática, Setor de Ciências Exatas, Universidade Federal do Paraná.
Orientador: Prof. Dr. Marcos S. Sunye

CURITIBA

2005

Agradecimentos

Gostaria de agradecer

A Deus, pelo grande presente: minha vida!

Aos meus pais, João e Alderi, e às minhas irmãs, Vivian e Simone, por fazerem parte da minha vida, e estarem sempre comigo. Mesmo à distância eles estiveram sempre me apoiando e dando forças para a realização do Mestrado e, sobretudo, para a conclusão desse trabalho.

Ao Fernando pela companhia e pelo carinho de sempre. Obrigada pela ajuda na escolha das imagens médicas e pela explicação sobre cada uma delas.

Ao professor Sunye por aceitar-me como sua aluna no Mestrado. Obrigada pela confiança e pela grande ajuda na realização desse trabalho.

Aos professores Luciano Silva e Hélio Pedrini, por ajudar-me nas questões referentes à área de Processamento de Imagens estudadas nesse trabalho.

Aos alunos Diego e Emanuel, que também foram cruciais para a minha familiarização com a área de Processamento de Imagens. Obrigada e sucesso para vocês!

Aos amigos do mestrado que foram muito importantes nessa etapa da minha vida e dos quais guardarei saudades: David e Pedro (obrigada pela ajuda com o Latex!), Diego, Bogdan, André, Cássio, José Augusto, Juliano, Nádia, Cláudio, Leslie, Paulo...

Resumo

Metadados são informações geralmente textuais vinculadas a recursos a fim de, sobretudo, promover meios para sua recuperação. Metadados são particularmente importantes para recursos visuais que se mantém sem nenhum texto, sendo, dessa forma, virtualmente irrecuperáveis. Bibliotecas Digitais utilizam metadados para fornecer meios para descrição e recuperação de seus recursos. Tais metadados são geralmente criados por profissionais, exigindo esforço, dedicação e tempo de pessoas experientes. Além disso, a descrição manual e textual de uma imagem pode trazer problemas quanto à subjetividade humana. Assim, torna-se necessário encontrar meios mais baratos, rápidos e eficientes para se criar metadados, e também para a indexação automática. A área de Recuperação de Imagens por Conteúdo surgiu para tentar superar estas dificuldades. Nesse caso, ao invés de serem anotadas por palavras-chaves, as imagens são indexadas pelo seu próprio conteúdo visual. Atualmente existem inúmeros algoritmos para extração de características visuais de imagens. Logo, esse trabalho tem como objetivo propor um mecanismo de geração automática de metadados de imagens das Bibliotecas Digitais através de características extraídas por algoritmos, facilitando e acelerando a criação desses metadados, e enriquecendo-os com novos elementos, a fim de prover uma melhor descrição das imagens. Imagens médicas serão utilizadas como exemplo. As características de conteúdo extraídas irão servir como indexadores mais precisos às imagens, complementando os tradicionais descritores qualitativos, e também permitirão que as imagens das Bibliotecas Digitais sejam recuperadas com base em similaridades, dada uma imagem exemplo. Logo, Provedores de Serviços serão capazes de realizar Buscas por Similaridade, uma importante ferramenta para aumentar a precisão das buscas às imagens.

Abstract

Metadata is generally textual information attached to a resource to aid its identification and retrieval. Metadata is particularly important for visual resources that might otherwise stand alone without any text, and therefore be virtually irretrievable. Digital Libraries of images use metadata to describe and improve resource retrieval. Experienced professionals, demanding effort, dedication and time, generally create metadata. Besides, the manual and textual description can create problems related to the human subjectivity. In the meantime, the generation, manipulation and retrieval of metadata automatically is cheaper, faster and efficient. So, it's necessary find conditions to automate the metadata creation and promote automatic indexing. The Content-based Image Retrieval area arose to try reducing the difficulties of image metadata creation. In this case, instead of annotate by keywords, the images are indexed by their own visual content. Recently several automated image-processing algorithms have been developed to image content extraction. So, the objective of this work is to propose a mechanism to automate the digital library image metadata creation through content features extracted by algorithms, accelerating and making easy the metadata creation, and enriching its with new elements, to provide a better description of the images. Medical images will be used as example. The extracted content features will serve as more precise image content indexes, complementing the traditional qualitative descriptions, and will allow the retrieval of images in digital libraries, based on similarities with a given image. Thus, Services Providers will be able to provide searches based on similarities, an important tool to increase the search of the images accuracy.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Todos os recursos precisam ter informações vinculadas a eles simplesmente para serem encontrados. Se esses recursos são imagens, tão importante quanto à forma que são apresentados, é o modo como são descritos, indexados e interpretados. A adição de informações, desde que de forma condizente e consistente, fornece meios para recuperação e promove maior visibilidade a uma imagem. Tais informações são chamadas metadados. Metadados são usualmente descritos como "dados sobre dados". Em um ambiente como a *Web*, eles são entendidos como os dados que podem auxiliar na organização, identificação, descrição, localização e recuperação de documentos eletrônicos e não eletrônicos. Bibliotecas Digitais utilizam metadados para, sobretudo, fornecer meios para descrição e recuperação de seus recursos.

Metadados são geralmente criados por profissionais experientes de forma manual e textual, exigindo muito trabalho dessas pessoas. No caso de imagens, duas dificuldades se apresentam: a necessidade de grande esforço manual para representá-las na forma de texto, e outra resultante do rico conteúdo das imagens e da subjetividade da percepção humana [18]. Isto é, pessoas diferentes podem perceber características diferentes pertencentes ao conteúdo de uma mesma imagem. A subjetividade da percepção humana e a imprecisão da anotação podem causar falhas graves nos processos de recuperação. Assim, torna-se necessário encontrar meios mais baratos, rápidos e eficientes para se criar metadados e prover indexação automática.

A área de Recuperação de Imagens por Conteúdo surgiu para tentar superar estas dificuldades. Nesse caso, ao invés de serem anotadas por palavras-chaves, as imagens são indexadas pelo seu próprio conteúdo visual, tais como, tamanho, cor, textura, forma, e outras características. Em uma típica recuperação de imagens por conteúdo, o usuário tem uma imagem na qual possui interesse e quer encontrar imagens similares no banco de dados.

Atualmente uma técnica automática e eficiente para a extração do conteúdo visual das imagens é o processamento das imagens por algoritmos. Tais algoritmos são capazes de extrair as características de uma imagem, as quais são armazenadas em vetores, chamados de Vetores de Características.

Esse trabalho propõe o uso de tais vetores de características como forma de enriquecer os metadados, e assim, dar início a um processo de automatização da criação dos metadados das imagens, o qual se tornará mais rápido, fácil e eficiente, uma vez que deixará de ser inteiramente manual.

Nesse caso, quando uma imagem é inserida em uma Biblioteca Digital, um algoritmo processa tal imagem e extrai determinada característica, a qual é armazenada em um vetor. Em seguida, um aplicativo poderá ler esse vetor de características e inserí-lo no documento que representa os metadados da imagem (geralmente um documento XML).

As Bibliotecas Digitais são uma das mais recentes novidades no mundo acadêmico informatizado e utilizam metadados para a disseminação de seus recursos. Elas são organizações que fornecem meios para selecionar, estruturar, oferecer acesso intelectual, interpretar e distribuir trabalhos digitais, de modo que estejam prontamente e economicamente disponíveis para comunidades. Em muitas áreas, no entanto, as imagens são efetivamente muito necessárias, como por exemplo, na medicina. Nesse caso, devido à importância das imagens médicas para a área da saúde é que está surgindo a idéia de Bibliotecas Digitais de Imagens Médicas, as quais ampliam o acesso às imagens como meio de pesquisa, beneficiando profissionais e estudantes que utilizam-nas para ensinar, aprender ou diagnosticar. Dessa forma, imagens médicas serão usadas como exemplo nesse trabalho.

Entretanto, para que os recursos das Bibliotecas Digitais possam ser efetivamente recuperados e compartilhados entre instituições e pessoas, é necessário estabelecer um comum acordo na adoção e uso de padrões de metadados. Metadados adicionados arbitrariamente a um recurso sem nenhum método estabelecido, não serão interoperáveis com outras instituições, podendo, conseqüentemente, ser de difícil localização e utilização [9]. Assim, a aplicação de metadados é controlada pelo uso de esquemas ou especificações, os quais consistem de elementos definidos para tipos específicos de informações. O Dublin Core [20] é o padrão de metadados mais utilizado para descrever os recursos da *Web*, e logo, das Bibliotecas Digitais.

Bibliotecas Digitais mantêm interoperabilidade entre si através da Iniciativa *Open Archives* (OAI) [24], a qual permite que os metadados das bibliotecas sejam compartilhados entre elas. A chave da OAI é o seu protocolo chamado OAI *Protocol for Metadata Harvesting* (OAI-PMH) [10], cujo propósito é o de coletar os metadados dos Provedores de Dados (por exemplo, Bibliotecas Digitais) e deixá-los disponíveis aos Provedores de Serviços. Os Provedores de Serviços são responsáveis por colher e unir em um só lugar os metadados de vários Provedores de Dados, e fornecer aos usuários mecanismos de buscas a esses metadados.

Uma vez que os metadados das imagens das Bibliotecas Digitais colhidos pelos Provedores de Serviços possuirão informações sobre o conteúdo visual dessas imagens, esse trabalho também sugere uma nova opção de busca a esses metadados: a Busca por Similaridade. Logo, além das buscas baseadas em informações textuais e palavras-chaves, as Buscas por Conteúdo também poderão ser usadas nos Provedores de Serviços.

Uma Busca por Similaridade envolve uma consulta com uma imagem exemplo, onde o usuário possui uma imagem e quer recuperar imagens similares. Sendo assim, o usuário pode acessar um Provedor de Serviços e entrar com uma imagem, a qual deverá ser processada por algoritmos a fim de ser extraído

seu conteúdo visual, na forma de Vetores de Características. O Provedor de Serviços, então, através de um sistema de cálculos de similaridade, compara o vetor de características da imagem de entrada com os vetores de características presentes nos metadados disponíveis, para então, selecionar as imagens similares. Questões referentes aos cálculos de similaridade por sistemas de recuperação de imagens por conteúdo, não fazem parte do escopo desse trabalho.

A Figura 1.1 ilustra a colheita de metadados por Provedores de Serviços através do protocolo OAI-PMH, a indexação automática e a busca por conteúdo propostas nesse trabalho:

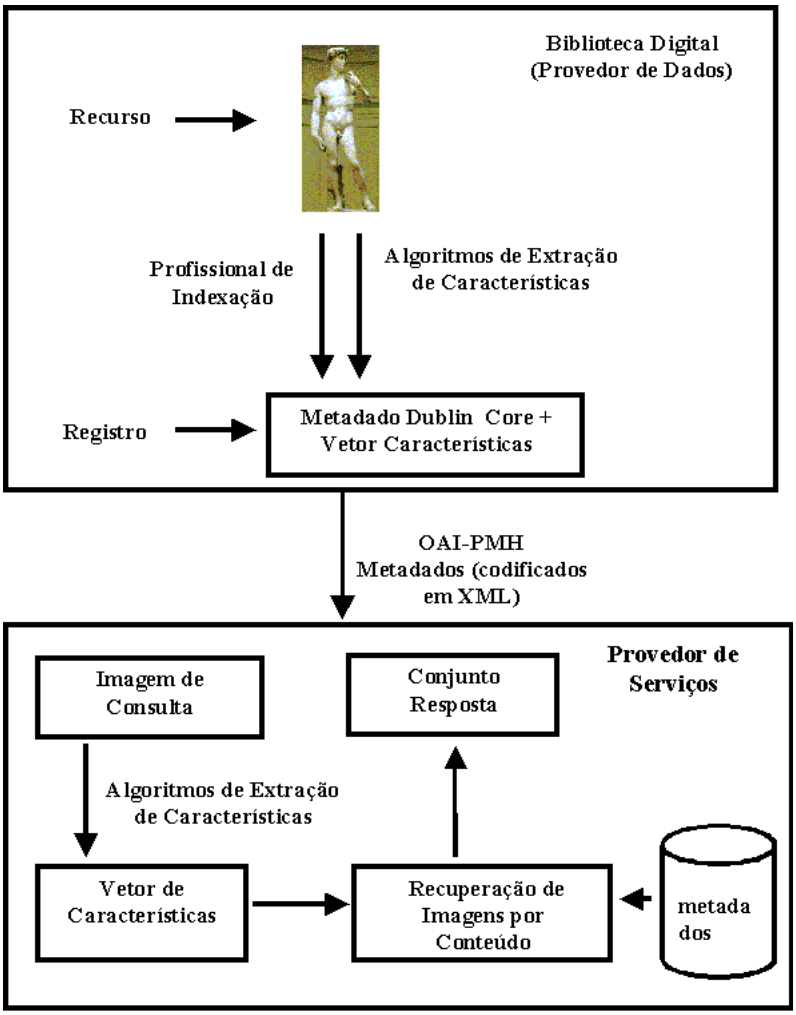

Figura 1.1: Provedores de Dados x Provedores de Serviços com Buscas por Conteúdo

Por essa figura, vê-se que os metadados das imagens das Bibliotecas Digitais podem ser criados manualmente por profissionais e automaticamente por algoritmos. Em seguida são colhidos pelos Provedores de Serviços através do protocolo OAI-PMH, armazenados e utilizados para buscas. Buscas por Conteúdos são possíveis. Nesse caso, o usuário entra com uma imagem exemplo para encontrar imagens similares. Tal imagem será processada por algoritmos que extrairão seu vetor de características, o qual será comparado através de um Sistema de Recuperação por Conteúdo com os metadados disponíveis, para que seja encontrado um conjunto resposta de imagens.

Tanto os Provedores de Dados quanto os Provedores de Serviços devem dar suporte ao formato de metadados baseado no Dublin Core, que é o formato usado pelo protocolo OAI-PMH para prover um nível mínimo de interoperabilidade entre seus participantes. Entretando, novos formatos podem ser propostos e validados junto à Iniciativa. Dessa forma, com a proposta desse trabalho de gerar novos elementos de metadados através das características extraídas por algoritmos, conseqüentemente está sendo proposta também a criação de um novo formato de metadados para as imagens das Bibliotecas Digitais. Logo, ainda serão estudadas nesse trabalho a iniciativa OAI e o protocolo OAI-PMH para prover interoperabilidade entre Bibliotecas Digitais, o formato de metadados utilizado pelo protocolo e as tarefas necessárias para a criação e validação de novos formatos de metadados junto à Iniciativa.

É importante ressaltar que a abordagem de geração automática dos metadados das imagens e buscas por conteúdo a esses metadados, proposta nesse trabalho, vale também para outros tipos de recursos. Além de imagens, Bibliotecas Digitais possuem outros tipos de mídia, tais como, vídeo e som. Esses recursos também podem ter seus conteúdos extraídos automaticamente por algoritmos e inseridos nos metadados. Dessa forma, as idéias a serem discutidas nesse trabalho com relação à imagens podem ser adaptadas para vídeos, por exemplo. Esse trabalho teve como estudo de caso as imagens devido à atual relevância desses recursos em várias aplicações, entre elas, projetos desenvolvidos no Departamento de Informática da UFPR.

1.1 Objetivos

Esse trabalho tem como objetivo propor um mecanismo para a automatização do processo de criação de metadados das imagens das Bibliotecas Digitais.

Para alcançar tal objetivo, será utilizada a técnica de extração automática de características visuais de imagens por meio de algoritmos. Na análise de imagens, as etapas responsáveis por extrair características são a Segmentação e a Extração, essa última também conhecida como Descrição. A ferramenta utilizada para o processamento das imagens será a CVIPtools [5], a qual implementa vários algoritmos importantes para o processamento de imagens. Através dessa ferramenta, as imagens serão segmentadadas e serão extraídas algumas características (descritores) da imagem. Para exemplificar o método proposto, serão extraídos descritores de forma e textura da imagem. Ainda como exemplo, serão utilizadas imagens médicas de vários tipos, tais como: raio-X, tomografia e ressonância magnética.

Quando as imagens são processadas por essa ferramenta, seu vetor de característica é gerado em um arquivo texto. Assim, para que determinada característica presente nesse vetor seja um novo elemento dos metadados da imagem (metadados das Bibliotecas Digitais são geralmente armazenados em documentos XML), um aplicativo é necessário para fazer a leitura desse vetor e inserí-lo no documento XML referente ao metadado. Esse aplicativo será desenvolvido na linguagem XSLT (*eXtensible Stylesheet Language for Transformations*) [45], que é uma linguagem capaz de processar e transformar documentos XML e cujos documentos são chamados de folhas de estilo. Dessa forma, a folha de estilo a ser desenvolvida irá transformar o documento XML que representa os metadados de uma imagem. Para tanto, ela irá ler os vetores de características das imagens e inserí-los nos documentos XML. Logo, tem-se que os metadados

das imagens possuirão elementos criados automaticamente, de uma forma rápida, eficiente e sem trabalho manual.

O segundo objetivo desse trabalho é propor que uma vez que esses metadados estejam enriquecidos com o conteúdo visual das imagens, a Recuperação de Imagens por Conteúdo estará habilitada aos sistemas que promovem buscas aos metadados das Bibliotecas Digitais, tais como os Provedores de Serviços. No caso dos Provedores de Serviços, o usuário poderá utilizar tanto a busca textual quanto a busca por conteúdo para obter seus resultados, uma vez que os metadados das imagens irão possuir tanto atributos textuais, já utilizados pelas Bibliotecas Digitais para descreverem seus recursos, quanto atributos visuais, extraídos por algoritmos.

Tanto os Provedores de Serviços quanto os Provedores de Dados devem estar de comum acordo quanto ao formato dos metadados, para que seja possível existir um nível mínimo de interoperabilidade, porém novos formatos podem ser propostos e validados junto à Iniciativa. O formato de metadados usado pela OAI é definido através da utilização de Esquemas XML. Uma vez que o trabalho propõe que os metadados das Bibliotecas Digitais possuam elementos adicionais que representem as características visuais dessas imagens, conseqüentemente um novo formato de metadados está sendo sugerido. Assim, o trabalho também apresentará as etapas necessárias para a validação junto à OAI de um novo formato de metadados, para que seja mantida a interoperabilidade entre seus participantes.

1.2 Organização

O Capítulo 2 irá introduzir o conceito de "metadados", a importância dos mesmos para descrição e recuperação de recursos e as vantagens da utilização deles. Também serão apresentados os padrões e especificações de metadados existentes, sobretudo, o Padrão Dublin Core.

O Capítulo 3 apresentará as Bibliotecas Digitais, suas principais finalidades e características. Também serão discutidas questões sobre a importância das Bibliotecas Digitais de Imagens e suas vantagens com relação aos Bancos de Dados de Imagens. Bibliotecas Digitais utilizam metadados para a descrição de seus recursos e são mais fáceis de serem utilizadas pelos seus usuários.

O Capítulo 4 discutirá sobre a Recuperação de Informações, sobretudo sobre os Sistemas de Recuperação de Imagens por Conteúdo e as Buscas de Imagens por Similaridade. Além disso, serão apresentados os principais métodos de consultas à informações, incluindo o método proposto pelo trabalho para a consulta das imagens das Bibliotecas Digitais.

O Capítulo 5 discutirá tópicos sobre o processamento de imagens. As etapas de segmentação e extração de características serão apresentadas. Ademais, serão discutidas as principais características visuais das imagens, entre elas as características usadas para a exemplificação da técnica proposta.

No Capítulo 6 serão apresentados conceitos sobre os Provedores de Serviços, que são sistemas capazes de promover buscas aos metadados das Bibliotecas Digitais. Provedores de Serviços e Provedores de Dados (tais como as Bibliotecas Digitais) são os dois tipos de participantes da Iniciativa de Arquivos Abertos (OAI). Assim, também será apresentada a OAI, como uma iniciativa com o intuito de prover meios para a interoperabilidade entre Bibliotecas Digitais, o protocolo OAI-PMH, desenvolvido pela

iniciativa para realizar a colheita dos metadados dos Provedores de Dados e o formato de metadados utilizado pelo protocolo para garantir um nível mínimo de interoperabilidade entre os participantes da OAI. Finalmente, como o trabalho propõe um novo formato de metadados para as imagens das Bibliotecas Digitais, esse capítulo apresentará questões quanto à criação de um novo esquema XML, para definir a nova estrutura desses metadados, e as tarefas necessárias para a validação desse novo esquema junto à Iniciativa.

O Capítulo 7 mostrará o desenvolvimento prático do trabalho. As tarefas realizadas foram: processamento das imagens pelos algoritmos de extração de características escolhidos, desenvolvimento da folha de estilo que irá adicionar os atributos visuais nos metadados das imagens e processamento desses metadados pela folha de estilo.

No Capítulo 8 será apresentada a conclusão do trabalho, discutindo os objetivos propostos, os resultados alcançados e os meios utilizados para alcançá-los. Ainda serão propostos tópicos para trabalhos futuros que possam dar continuidade à idéia proposta nesse trabalho.

Capítulo 2

Metadados

Metadados são informações vinculadas a um recurso, geralmente na forma de textos e palavras-chaves. Estas informações podem ser relativamente diretas, tais como nome do autor, data de criação, assunto, mas também, podem ser mais complexas e mais difíceis de serem definidas, tal como o consenso da opinião de várias pessoas sobre um mesmo livro [9].

A informação contida nos metadados é "pesquisável" e, conseqüentemente, auxilia na identificação e recuperação dos recursos. Metadados não só ajudam na descoberta dos recursos mas também no entendimento da natureza do que foi encontrado. Eles também ajudam o usuário a avaliar o recurso, fazer um julgamento sobre o mesmo, compará-lo com outros recursos e avaliar sua adequação para determinado propósito.

Metadados são particularmente importantes para recursos visuais que se mantêm sem nenhum texto e que, conseqüentemente, são virtualmente irrecuperáveis. No caso das imagens, por exemplo, os usuários dependem das informações adicionadas a seus metadados para busca e recuperação eficientes. Mas os metadados não são só importantes para o usuário final: se apropriados, metadados técnicos e administrativos (tais como, informação sobre direitos autorais, informação sobre o processo de criação de imagens, formato do arquivo, resolução, etc) também auxiliam no gerenciamento, manutenção e preservação das coleções digitais.

Podem ser destacadas as seguintes vantagens na utilização e disponibilização de metadados: facilidade e maior precisão na recuperação das informações desejadas; estabelecimento de padrões de dados diante da heterogeneidade de informações contidas em rede, como por exemplo, na Internet; troca de informações entre aplicações e organizações; etc [12].

Existem algumas opiniões na forma de categorizar os tipos de metadados. Gilliland-Swetland [13] definiu cinco tipos básicos de metadados, que podem ser vistos na Tabela 2.1.

Este trabalho tratará dos metadados descritivos, responsáveis por apresentar os conteúdos visuais das imagens e ajudar na localização das mesmas.

Tabela 2.1: Tipos de Metadados

Tipo	**Definição**
Administrativo	Metadados usados no gerenciamento e administração das informações, por exemplo, direitos autorais, aquisição do recurso, permissões e outras informações usadas para gerenciar acessos.
Descritivo	Metadados usados para descrever ou identificar os recursos, por exemplo, assunto do conteúdo do recurso, vocabulários controlados, etc.
Preservação	Metadados relacionados ao gerenciamento da preservação dos recursos, como por exemplo, condições físicas do recurso, ações de preservação, etc.
Técnico	Metadados relacionados a formatos e estruturas, tais como, informações sobre a digitalização do recurso, compressão, etc.
Uso	Metadados relacionados ao nível e tipo de uso do recurso.

2.1 Metadados Descritivos

Metadado descritivo é geralmente a informação para descrever ou identificar um recurso. Os metadados descritivos são freqüentemente expressados como texto, podendo ser usados para descrever informações que estão em outros formatos, tais como imagens, sons, mapas e programas de computador.

O metadado descritivo é geralmente criado por profissionais especializados. No caso de catálogos bibliográficos e índices científicos, por exemplo, existe um grande investimento de trabalho de pessoas experientes há décadas e até mesmo, séculos. Este fato econômico é crucial para o entendimento da tendência atual. Hoje é necessário encontrar meios mais baratos e rápidos para se criar metadados, e também para indexação automática, através do uso de ferramentas que cheguem à especialidade humana.

2.2 Catálagos

Os catálogos são pequenos registros que fornecem informações sobre um objeto de uma biblioteca. A palavra “catálogo” é aplicada para registros que possuem uma estrutura consistente, organizada de acordo com regras sistemáticas.

Catálogos de bibliotecas apresentam muitas funções, não somente a recuperação de informações. Alguns catálogos oferecem informações bibliográficas compreensivas que não podem ser derivadas diretamente dos objetos. Isto inclui informações sobre autores ou a procedência dos artefatos de um museu.

Para gerenciar coleções, os catálogos contêm informações administrativas, tais como permissões e outras informações para gerência de acesso.

Um catálogo pode combinar registros de todos os gêneros, mídias e formatos. Isto capacita os usuários de bibliotecas digitais a descobrirem vários tipos de material através de busca de registros textuais. Em bibliotecas convencionais, os materiais que são armazenados sobre enormes estantes são descritos por registros que podem estar contidos em um banco de dados *on-line*.

As informações, dentro de um catálogo, são divididas em "campos" e "subcampos" com *tags* para identificá-los. Em função do trabalho exigido para se criar e manter um registro de catálogo detalhado, em muitas bibliotecas os materiais são catalogados uma só vez e os registros são então distribuídos para outras as bibliotecas.

2.3 Serviços de Resumo e Indexação

Geralmente, os usuários da área científica querem informações sobre assuntos específicos. Por causa da sutileza da linguagem textual, a busca por assuntos é imprecisa, a menos que haja indexação da informação que descreve o assunto de cada objeto. A informação do assunto pode ser um resumo, palavras-chave, ou outras informações. Alguns serviços pedem aos autores que eles forneçam as palavras-chave ou um resumo do seu documento, mas isto conduz a grandes inconsistências. Por outro lado, ainda está longe de se ter indexadores profissionais capazes de indexar documentos de assuntos variados.

Uma abordagem bastante efetiva para tratar inconsistência, porém cara, é o uso de um "vocabulário controlado". Nesse caso, vários termos pré-definidos podem ser usados para descrever um conceito. O indexador dá uma lista de termos de assuntos aprovados e regras para aplicá-los. Essa abordagem requer indexadores treinados e usuários experientes. Isto porque os termos usados pelo usuário em uma busca devem ser consistentes com os termos designados pelo indexador.

2.4 Indexando Imagens

Quando uma coleção possui um assunto específico e uma comunidade bem definida de usuários, o processo de indexar e representar imagens é mais fácil, pois os possíveis propósitos para os quais as imagens serão utilizadas, podem ser antecipados. Termos apropriados, possivelmente de um vocabulário controlado, podem ser usados para facilitar a recuperação, especialmente se foi realizada alguma pesquisa inicial sobre os usuários e os termos por eles utilizados.

Quando uma coleção é multidisciplinar, ou seja, abrange vários assuntos, e possui uma comunidade diversa de usuários, a indexação é uma tarefa muito mais complexa, pois é impossível prever como usuários de diferentes áreas e com diferentes propósitos em mente para a utilização da imagem, irão procurar as imagens, e quais termos de busca eles utilizarão.

Por exemplo, a imagem mostrada na Figura 2.1 é um tecido de uma glândula humana, cuja utilização está voltada principalmente a médicos, enfermeiros e estudantes de medicina. Entretanto a imagem poder ser usada por um estudante de arte que esteja procurando uma inspiração para o design de um

trabalho. Dessa forma, observa-se que o conteúdo de uma imagem pode ter uma variedade de significados, dependendo das necessidades dos usuários.

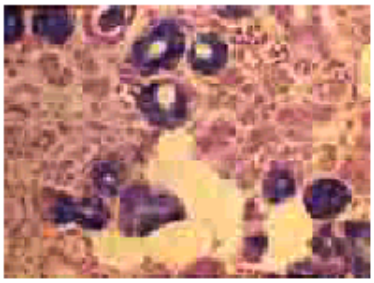

Figura 2.1: Tecido do Corpo Humano

A representação de imagens geralmente é um processo difícil. Uma imagem envolve conceitos relacionados à realidade a qual ela representa. Tais conceitos não são facilmente representados por palavras; nesse caso, são procuradas características visuais da imagem. Da mesma forma, a realidade nem sempre pode ser expressa por palavras.

A maneira na qual uma imagem vai ser representada terá conseqüências para sua facilidade ou não de ser recuperada por um usuário.

A representação de imagens é um problema se for realizada por pessoas não qualificadas, pois é um processo que requer tempo e custos. Isso leva a crer que é estritamente necessário que essas pessoas sejam especializadas no domínio na qual elas trabalham, mas somente isso não é a solução.

2.5 Padrões de Metadados, Esquemas e Especificações

Todos os recursos precisam ter informações vinculadas a eles para serem encontrados. Entretanto, para que os recursos sejam efetivamente recuperados e compartilhados entre instituições e pessoas, é necessário estabelecer um comum acordo na adoção e uso de padrões para a adição dessas informações. Metadados adicionados arbitrariamente a um recurso sem nenhum método estabelecido, não serão interoperáveis com outras instituições e, conseqüentemente, serão difíceis de serem localizados, e logo, usados [9].

Metadados interoperáveis descrevem objetos de uma mesma maneira, podendo ser compartilhados entre organizações, aumentar o escopo das coleções e possibilitar a implementação de sistemas de buscas entre essas coleções. Para um nível mínimo de interoperabilidade, esses metadados devem conter um certo número dos mesmos elementos de descrição.

Estudos sobre a semântica também contribuem para a construção de novos e interoperáveis padrões de metadados, que serão apropriados às Bibliotecas Digitais para possibilitar o intercâmbio de informações. Este intercâmbio permitirá a operação integrada de diferentes sistemas e bases de dados. Um exemplo é a Biblioteca Digital de Teses e Dissertações (TEDE) do IBICT [6], que adere ao padrão *Open Archives* de intercâmbio de metadados.

A aplicação dos metadados é controlada pelo uso de esquemas ou especificações. Tais esquemas ou especificações consistem de elementos definidos para tipos específicos de informações. Dessa forma,

elementos de metadados são componentes individuais que fazem parte de um esquema. Cada elemento irá conter um tipo particular ou categoria de informação, dependendo da sua definição. Por exemplo, a maioria dos metadados contém um elemento Autor, o qual informa o nome da pessoa que criou ou originou o objeto. Um esquema de metadados é geralmente comparado a um registro MARC [26], sendo esse último uma descrição estruturada de um recurso usada em catálogos de bibliotecas para indexar tal recurso com base em textos.

Metadados estão sendo usados de diversas formas por vários grupos. Como resultado, muitas especificações e padrões têm sido desenvolvidos. Elementos variam entre especificações, nem todos os esquemas possuem os mesmos elementos, diferenciando de acordo com as necessidades das comunidades que irão utilizá-los(por exemplo, museus, galerias de arte, fundações educacionais, comunidades médicas), onde a importância de tipos e categorias de elementos é distinta entre elas.

Como exemplo de diferentes padrões de metadados cuja utilização depende da finalidade associada tem-se: FGDC (*Federal Geographic Data Committee*) para descrição de dados geo-espaciais; MARC (*Machine Readable Catalogue*) para catalogação bibliográfica; IAFA/WHOIS++ (*Internet Anonymous Ftp Archive with Whois++ protocol*) para descrição do conteúdo e serviços disponíveis em arquivos ftp (*file transfer protocol*); TEI (*Text Encoding Initiative*) para representação de materiais textuais na forma eletrônica; DC (*Dublin Core*) para desrição de recursos na *Web*; SAIF (*Spatial Archive and Interchange Format*) para compartilhamento de dados espaciais e espaço-temporais [12].

Esse trabalho estudará o padrão de metadados Dublin Core, que é o padrão mais utilizado para a descrição de recursos da *Web*, e conseqüentemente, para os recursos das Bibliotecas Digitais.

2.5.1 Dublin Core Metadata Initiative (DCMI)

O esquema que tem sido mais amplamente usado nos últimos anos para descrever recursos da *Web* é o DCMES (*Dublin Core Metadata Element Set*). O DCMES [37] compreende um simples conjunto de quinze elementos genéricos aplicáveis a uma variedade de tipos de objetos digitais. Ele tem sido adaptado por inúmeras comunidades científicas para acomodar as necessidades próprias de tais comunidades, e tem como objetivo principal ser simples, facilitando a aplicação por parte do usuário na descrição dos seus recursos.

A Iniciativa de Metadados Dublin Core (DCMI) é uma iniciativa que surgiu em Dublin, Ohio (EUA) em 1995 e que se dedica a promover a adoção de padrões de interoperabilidade em metadados.

O Dublin Core é usado para completar métodos existentes para buscar e indexar metadados na *Web*. As primeiras discussões que ocorreram concentraram-se, principalmente, em criar metadados para recursos eletrônicos. Entretanto, a partir destas discussões, o consenso entre a comunidade Dublin Core é que sistemas de descoberta de recursos podem e devem ser usados para descrever objetos físicos digitais e não digitais. A maioria dos participantes da DCMI está envolvida amplamente no arquivamento ou catalogação de projetos que exigem o uso do Dublin Core para habilitar grandes coleções de objetos para que eles sejam agrupados, nomeados, classificados e indexados de uma forma útil.

Os quinze elementos do DCMES podem ser vistos na Tabela 2.2.

Tabela 2.2: Quinze Elementos Dublin Core

Tipo de Elemento Dublin Core	**Definição**
Title	Nome dado ao recurso.
Creator	Nome da pessoa ou organização responsável pela criação do conteúdo intelectual do recurso.
Subject	Tópico relacionado ao conteúdo do recurso. Indicado por palavras-chave e por códigos de esquemas de classificação.
Description	Descrição textual do conteúdo do recurso, ou uma referência para essa descrição, por exemplo, um resumo ou abstract.
Publisher	A entidade ou pessoa responsável por tornar o recurso disponível.
Date	A data associada com a criação ou disponibilidade do recurso, no formato ISO 8601 (ano-mês-dia).
Contributor	Uma pessoa ou organização não especificada no elemento creator, que tenha contribuído ao conteúdo intelectual do recurso.
Type	A categoria ou gênero do recurso, tal como um texto ou uma imagem.
Format	O formato dos dados do recurso, usado para identificar possíveis softwares ou hardwares necessários para exibir ou operar o recurso.
Identifier	Uma string ou números usados para identificar recursos de forma única, por exemplo, URL ou ISBN.
Source	Informação sobre um recurso secundário do qual o presente recurso foi derivado.
Language	A linguagem utilizada para expressar o conteúdo intelectual do recurso.
Relation	Indica o relacionamento do presente recurso com outros recursos.
Coverage	Características da extensão ou do escopo do recurso, tais como: cobertura espacial, temporal, jurisdição.
Rights	Indica as referências ou direitos de propriedade sobre o recurso.

Usando o Dublin Core para estruturar as informações do documento desse Trabalho, elas ficariam da

seguinte forma:

Tabela 2.3: Metadados Dublin Core desse Documento

Tipo de Elemento Dublin Core	**Definição**
Title	Um Mecanismo para Automatizar a Criação dos Metadados das Imagens de Bibliotecas Digitais e Prover Buscas por Conteúdo
Creator	Beatriz do Carmo Langiano
Contributor	Marcos Sunye
Subject	Metadados, Bibliotecas Digitais, Indexação de Imagens, Algoritmos de Extração de Conteúdo, Provedores de Serviços.
Date	2005-05-25
Type	Text
Identifier	http://www.inf.ufpr.br/ beatriz/dissertacao.pdf
Rights	Departamento de Informática - Universidade Federal do Paraná.

Alguns elementos possuem convenções para a maneira pela qual seus valores são introduzidos, como os elementos Date e Type. O elemento Type possui uma lista (vocabulário controlado) chamada DCMI Type Vocabulary [42], usada para determinar a natureza do conteúdo de um recurso. Atualmente os termos dessa lista são:

- Collection
- Dataset
- Event
- Image
- Interactive Resource
- Service
- Software
- Sound
- Text

Além dos quinze elementos Dublin Core citados na Tabela 2.2, existem também os chamados "qualificadores Dublin Core". Esses qualificadores servem para refinar o significado de um recurso. Eles permitem às aplicações aumentarem a precisão dos metadados. Por outro lado, introduzem complexidade que

poderia prejudicar a compatibilidade dos metadados com outros softwares de aplicações Dublin Core. O "date" é um exemplo de elemento que possui qualificadores para refinar e identificar um tipo particular de data (por exemplo, data da última modificação, data de publicação, etc). O uso de vocabulários controlados, tal como o DCMI Type Vocabulary citado acima, é outro método usado para promover qualidade ao significado dos recursos.

O conjunto dos quinze elementos de metadados do Dublin Core é também às vezes chamado de "*Dublin Core Unqualified*", para ressaltar que nenhum qualificador é usado.

Análise do Padrão Dublin Core

Este padrão tem como objetivo principal, ser de extrema simplicidade, facilitando a aplicação por parte do usuário na descrição dos recursos *Web*. Quando da aplicação do padrão em imagens digitais, verifica-se que os descritores são muito genéricos (como por exemplo, o descritor "tipo de recurso") ou muito restritivos (como por exemplo, a concatenação do "assunto" e das "palavras-chaves"). Assim, constata-se que a aplicação direta do padrão para a descrição de documentos eletrônicos de uma imagem digital é insuficiente, uma vez que nesse tipo de aplicação o foco não está somente na descoberta do recurso, mas também na sua descrição sendo, nesse caso, a aplicação do padrão insuficiente para representar o conteúdo em poucos descritores.

Capítulo 3

Bibliotecas Digitais

A Internet é uma rica infra-estrutura para o avanço tecnológico que tem influenciado o surgimento de novas técnicas e paradigmas para recuperação de informações. A criação de Bibliotecas Digitais tem sido uma das formas encontradas para a organização e democratização do acesso às informações.

Waters [43] define em seu artigo as Bibliotecas Digitais como: "Organizações que fornecem serviços, incluindo uma equipe especializada para selecionar, estruturar, acessar, interpretar, distribuir e preservar coleções de trabalhos digitais de modo que estejam prontamente e economicamente disponíveis para as comunidades".

As Bibliotecas Digitais podem possuir qualquer informação que possa ser codificada como uma seqüência de bits. Essas informações estão disponíveis eletronicamente ao público por meios apropriados tais como textos, imagens, vídeos ou voz.

Dessa forma, as bibliotecas digitais podem ser vistas como coleções distribuídas de objetos digitais, que cobrem várias áreas de interesse humano, tais como arte, música, medicina, ciência, filmes, livros, literatura, jornais, etc. Os objetos digitais das coleções apresentam diversas características interessantes: podem ser copiados indefinidamente sem perder qualidade, não desgastam com o manuseio e com o tempo, ocupam pouco espaço físico ao serem armazenados, além de poderem ser distribuídos pela Internet e recuperados remotamente.

As bibliotecas digitais, pelas características de seu acervo e mídias, são ambientes propícios às atividades derivadas da automatização tecnológica e manipulação digital dos documentos. Uma das linhas de pesquisa consideradas diz respeito às possibilidades de indexação automática dos documentos ao entrarem para bases de dados das bibliotecas digitais, a partir de diversas estratégias. Estas estratégias permitirão às bibliotecas digitais o oferecimento de novas formas de buscas semânticas sobre os documentos, somando possibilidades aos processos tradicionais de recuperação de informações e ampliando o processo de comunicação da informação científica.

Esse trabalho propõe um modelo de indexação automática das imagens das Bibliotecas Digitais. Essa

automatização se dará através do uso de algoritmos que serão responsáveis pela extração de características visuais das imagens. Feito isso, essas características serão lidas por um aplicativo e automaticamente inseridas no metadado da imagem correspondente. Dessa forma uma nova forma de busca estará disponível: além da busca textual (os metadados das imagens possuem campos descritos textualmente), a busca por conteúdo será uma nova opção.

3.1 Características das Bibliotecas Digitais

Abaixo se seguem algumas características que ajudarão a explicar mais claramente o significado das Bibliotecas Digitais:

- Fornecer serviços: As bibliotecas digitais são organizações que empregam uma variedade de serviços, e que não necessitam ser organizadas no modelo das bibliotecas convencionais. Embora as necessidades que as bibliotecas digitais requerem sejam similares àquelas dentro das bibliotecas convencionais, elas são, muitas vezes, de tipos diferentes. Por o exemplo, para o armazenamento e recuperação, as bibliotecas digitais dependem quase exclusivamente do computador e dos sistemas de rede.

- Coleções de trabalhos digitais, onde as distinções entre as bibliotecas geralmente estão no foco do tema que define as coleções (por exemplo, medicina, arte, ciência, música, e outros), ou nas comunidades interessadas nos materiais coletados (por exemplo, pesquisa, faculdade, público).

- Úteis por uma comunidade ou por um conjunto de comunidades definidas: As bibliotecas em geral, e as bibliotecas digitais particularmente, são organizações de serviços. As necessidades e os interesses das comunidades que elas servem determinarão a trajetória do desenvolvimento das bibliotecas digitais, incluindo o investimento feito no conteúdo e na tecnologia. A maioria das bibliotecas digitais é dedicada ao suporte à educação e pesquisa, e justificam seu investimento em desenvolvimentos digitais como um meio poderoso de realizar os objetivos institucionais das comunidades acadêmicas as quais elas servem.

3.2 Metadados e Interoperabilidade entre Bibliotecas Digitais

Itens de uma biblioteca digital são chamados de objetos digitais. Eles são armazenados em repositórios e identificados por "handles". As informações armazenadas em um objeto digital são chamadas de conteúdo, que são divididos entre Dados e Metadados. "Dado" é um termo genérico usado para descrever informações que estão codificadas na forma digital. Metadados são os "dados" sobre esses dados.

Uma questão importante quanto às Bibliotecas Digitais é a disponibilização das suas informações. Universidades e Centros de Pesquisas são grandes geradores de conhecimento. Com as bibliotecas tradicionais, em um passado recente, a disponibilização do conteúdo estava sujeita a uma infraestrutura básica e, muitas vezes, de difícil acesso.

Nos últimos anos, com o avanço da Internet, várias Bibliotecas Digitais começaram a surgir com a finalidade principal de expor a produção de teses e dissertações. Atualmente, autores, publicadores e

outros provedores de conteúdos estão cada vez mais interessados em participarem das bibliotecas digitais, tornando seus conteúdos disponíveis a vários tipos de públicos.

Entretanto, a falta de padrões para disponibilização e pesquisa de informações científicas na Internet levou à criação da Iniciativa *Open Archives* (OAI) e ao desenvolvimento de um protocolo com o intuito de oferecer simplicidade e eficiência na tarefa de unificar as consultas à base de dados científicas/acadêmicas. Com os recursos oferecidos pela OAI é possível melhorar significativamente a precisão das consultas eletrônicas e reduzir o tempo de procura, graças ao compartilhamento de informações (metadados) entre os participantes da Iniciativa.

Mais informações sobre a Iniciativa *Open Archives* estarão disponíveis no Capítulo 6.

3.3 Bibliotecas Digitais de Imagens

Antes da criação das bibliotecas digitais, as imagens eram armazenadas em banco de dados de imagens. Estes bancos eram de grande utilidade, mas causavam alguns problemas, sobretudo com relação à semântica dessas imagens e ao acesso as mesmas.

A forma estruturada de armazenamento dos dados em banco de dados não permite trabalhar adequadamente com objetos multimídia. Um problema básico que ainda incomoda a quem precisa gerenciar informações que incluam, além de texto, imagens, áudio e vídeo, se refere à manipulação desses objetos, uma vez que, além do simples armazenamento, é importante conseguir disponibilizar informações de conteúdo semântico destes objetos, de forma a facilitar sua recuperação [12]. Freqüentemente, as imagens não são descritas de forma adequada.

Quanto ao acesso às imagens, falta ainda um meio mais natural de consultar e acessar esses bancos. Utilizar esses bancos normalmente exige um conhecimento aprimorado de técnicas de buscas eletrônicas e muitas vezes do próprio assunto pesquisado, limitando sua utilização a profissionais experientes.

Essa dificuldade de acesso às imagens, entre outras coisas, levou ao desenvolvimento das Bibliotecas Digitais de Imagens. Essas bibliotecas possuem descritores associados aos dados contidos nas imagens, conhecidos como metadados. Os metadados incluem elementos de descrição do conteúdo dos dados e qualquer informação relevante para a recuperação dos seus conteúdos.

Bibliotecas Digitais de Imagens Médicas ampliam os meios de acesso às imagens, sendo responsáveis pelo armazenamento e recuperação dessas imagens. Elas beneficiam profissionais e estudantes que utilizam-nas para diagnosticar, ensinar ou aprender, possibilitando que, médicos, alunos e professores passem a ter acesso a informações que não podiam ser representadas sob a forma de texto.

Capítulo 4

Recuperação de Informações e Buscas por Conteúdo

O termo Recuperação de Informação, ou IR (*Information Retrieval*) descreve o processo através do qual um usuário converte uma consulta em uma coleção de referências úteis [15]. O autor dessa definição, Calvin Moores, refere-se à informação textual. Em um sistema de IR, um documento é representado como uma coleção de características, tais como palavras-chaves, citações, referências bibliográficas [27]. O usuário especifica a informação de que precisa através de uma consulta. Dada a consulta do usuário e uma coleção de documentos, o sistema busca os documentos mais relevantes à consulta.

Este conceito pode ser estendido para recuperação de informações visuais. O interesse em adaptar este conceito às imagens surgiu devido à velocidade dos avanços na área computacional. A facilidade de captura e compreensão de imagens digitais tem produzido uma quantidade gigantesca de informação visual *on-line* [1], e conseqüentemente, aumentado em muito a popularidade das Bibliotecas Digitais de Imagens. Com isso, nos últimos anos observou-se um rápido aumento no tamanho das coleções de imagens digitais.

Tradicionalmente, a indexação destas imagens é feita de forma manual e textual. A idéia é fazer anotações em forma de texto sobre as imagens e usar gerenciadores de banco de dados convencionais para fazer a recuperação. Devido a isso, uma abordagem muito utilizada para recuperação de imagens é a baseada em texto. Esta primeira solução, deve-se em parte, à evolução da recuperação no campo textual, com o desenvolvimento de sistemas capazes de recuperar a informação de forma eficiente, como o AltaVista, o Google e o Yahoo.

Logo, duas dificuldades se apresentaram: a necessidade de muito esforço manual para representar imagens em forma de textos; e outra que resulta do rico conteúdo das imagens e da subjetividade da percepção humana [18]. Isto é, pessoas diferentes podem perceber características diferentes pertencentes ao conteúdo de uma mesma imagem. A subjetividade da percepção e a imprecisão da anotação podem

causar falhas graves nos processos de recuperação [30].

A área de Recuperação de Imagens por Conteúdo surgiu para tentar superar estas dificuldades. Isto é, ao invés de serem anotadas somente por palavras-chaves, as imagens seriam indexadas também pelo seu próprio conteúdo visual. Assim, o problema da recuperação de imagens é um aspecto particular do tratamento geral de recuperação de informação.

Portanto, a análise do conteúdo e a indexação baseada no conteúdo visual têm sido áreas de muitos estudos, pois é evidente a necessidade de indexar e processar imagens de tal maneira que se consiga consulta e recuperação eficientes.

4.1 Recuperação de Imagens por Conteúdo e Indexação de Imagens

De acordo com a técnica de Recuperação de Imagens baseada em Conteúdo, no momento da inserção da imagem no sistema, são extraídas as principais características visuais da imagem (tais como cor, textura, contorno, formas, curvatura, etc.). Essas características formam o Vetor de Características (*feature vector*) de uma imagem, que é armazenado junto com ela em uma base de dados. No momento de uma consulta, o vetor de características que representa a imagem de consulta é computado e o SGBD compara este vetor com cada vetor armazenado na base de dados. Esta operação pode ter um custo de processamento elevado se o conjunto de dados a ser pesquisado for grande.

A necessidade de otimizar a recuperação dos dados deu origem a diversas pesquisas que resultaram em uma grande variedade de "Estruturas de Indexação". No caso dos Sistemas de Recuperação de Imagens por Conteúdo (SRICs), a função da estrutura de indexação é prover acesso rápido aos objetos de dados, tornando esses sistemas escaláveis para grandes coleções de imagens. Cada método de acesso procura montar um mecanismo que represente de forma simplificada as informações que endereçam e definem uma série de operadores relacionais para desempenhar operações de buscas na estrutura criada.

Porém conforme afirma Petrakis [31], "*A recuperação de imagens não é um processo exato (imagens raramente são idênticas)*". Dessa forma, o mecanismo de busca de um SRIC não pode ser muito rigoroso a ponto de não recuperar *boas-candidatas*, nem tampouco ser flexível demais que recupere muitas *falsas-candidatas*. O ideal é que seja adaptável ao domínio da aplicação. Ou seja, a estrutura de indexação deve prover um mecanismo de filtragem que elimine de forma rápida as *falsas-candidatas*, reduzindo assim o espaço de dados a ser pesquisado.

Nesse caso, deve existir um conceito de similaridade implementado no banco de dados do SRIC: uma imagem é similar à outra se os valores de determinadas características são próximos. Uma recuperação eficiente retornaria como resposta a uma consulta todas as imagens cujo vetor de características se assemelhe ao vetor de características da imagem de consulta.

Resumidamente, a idéia geral de um SRIC é criar para cada imagem um vetor de características que seja uma representação o mais próximo possível, da imagem armazenada e implementar uma estrutura de indexação e operadores que permitam realizar pesquisas eficientes sobre esses vetores.

A representação de uma imagem é feita por meio de um conjunto de características extraídas por um processo automático composto por um grupo de algoritmos. Este mesmo processo é aplicado sobre uma

certa quantidade de imagens, representando-as por vetores de características de mesmas dimensões.

Geralmente uma única característica não basta para identificar e garantir a unicidade de uma imagem em um banco de dados de imagem. Desta forma, para inserir uma imagem em um SRIC, deve-se armazenar a imagem física e tantos atributos quanto forem necessários para prover uma recuperação eficiente desta imagem.

Os SRICs utilizam tanto atributos de baixo nível de uma imagem (cor, textura, forma e outras características), quanto informações semânticas na forma de texto (palavras-chaves e anotações).

Em uma típica recuperação de imagens por conteúdo, o usuário tem uma imagem na qual possui interesse e quer encontrar imagens similares. O cenário aqui envolve uma consulta com uma imagem exemplo, onde o usuário entra com uma imagem e quer recuperar imagens que tenham alguma semelhança com a imagem de entrada. O primeiro problema é encontrar características adequadas para a representação da imagem. Em seguida, deve-se usar uma medida eficiente para estabelecer similaridade entre duas imagens. Questões referentes às técnicas de medidas de similaridade entre imagens não fazem parte do escopo desse trabalho.

4.2 Métodos de Consulta em Sistemas de Recuperação de Imagens

Nos últimos anos a recuperação de imagens por conteúdo se tornou uma crescente área de estudo. Muitos sistemas de recuperação de imagens foram criados. A maioria desses sistemas suporta métodos de consulta tais como:

- *Browsing*
- Consulta por Imagem Exemplo
- Consulta por Esboço
- Consulta por Texto
- Consulta por Características

4.2.1 Browsing

É uma técnica interativa de recuperação de informações, que usa as capacidades cognitivas humanas para evitar os problemas de interação homem-máquina (os humanos são melhores para reconhecer a informação desejada do que para descrevê-la). Nessa técnica, a formulação de uma consulta ao sistema não requer que o usuário tenha conhecimento de uma linguagem de consulta ou da arquitetura do banco de dados. O navegador possui apenas botões que indicam ao usuário o caminho a seguir.

Este método é adequado quando a informação está mal-definida, por ajudar o usuário a determinar a informação necessária, ou quando o usuário quer ter uma impressão sobre o conteúdo de uma coleção de dados, onde o navegador oferece uma visão geral dos dados da coleção.

Um exemplo de aplicação pode ser observado quando um pesquisador navega em um sistema de um museu para observar as obras disponíveis para estudo. A princípio, o estudante pode não ter em mente

uma obra ou um artista específico. Seu objetivo é apenas navegar pelo sistema e apreciar o acervo do museu [30].

4.2.2 Consulta por Imagem Exemplo

Consulta por Imagem Exemplo, como o próprio nome diz, é uma consulta na qual o usuário tem interesse em encontrar imagens similares a uma imagem que ele tem como exemplo.

Um dos maiores problemas da recuperação de imagens por conteúdo é que sistemas tradicionais de busca trabalham apenas com o conceito de igualdade, o que não seria útil no caso da busca por imagens. A abordagem consulta por imagem exemplo não lida com igualdade, mas sim com operadores de similaridade.

Esta abordagem é muito utilizada em sistemas que usam as características cor e textura e em sistemas onde é comun o usuário possuir uma imagem para a consulta. Por exemplo, um profissional na área médica pode ter interesse em encontrar uma imagem de mamografia. Como esse tipo de imagem é comum nessa área, é possível que o profissional possua um exemplo disponível, e deseje procurar uma outra com características semelhantes.

4.2.3 Consulta por Esboço

Essa abordagem assemelha-se ao caso ao anterior, já que o usuário deve fornecer como entrada um esboço da informação que precisa. O sistema pode oferecer uma ferramenta interativa que permita que o usuário faça um desenho sobre o assunto que procura ou oferecer partes das imagens para o usuário compor um esboço a partir de fragmentos de imagens.

Esta abordagem é muito utilizada para recuperar imagens que contenham objetos com formas similares ao esboço. Um exemplo de aplicação é a busca de fotos de pessoas baseada em retratos falados [30].

4.2.4 Consulta por Texto

Uma maneira intuitiva de descrever uma imagem é usando palavras. Estas abstrações são chamadas descrições ou anotações. Consulta Direta por Descrições é o método de recuperação no qual uma consulta por imagens consiste apenas de valores de características especificados pelo usuário. Estes valores são associados às imagens através do conhecimento humano no momento da sua inserção. Assim, a consulta se referirá às descrições e o resultado será um conjunto de imagens associado às descrições, sendo necessário o uso de um método de indexação.

Já existem sistemas de recuperação de imagens automáticos que utilizam a semântica da imagem e processamento de linguagem natural, onde a recuperação é feita com base nas descrições. Entretanto existem falhas, pois esses sistemas não são capazes de recuperar totalmente a semântica de uma imagem.

Assim, o trabalho de associar palavras-chaves a imagens pode ser superado, mas a ambigüidade das descrições continua sendo um problema. Um exemplo disso é quando uma imagem está associada a uma palavra-chave em um contexto e a outra em um outro momento.

Para a indexação por descrições geralmente uma equipe de pessoas é contratada. Nesse caso, mesmo que a equipe seja especializada, pode acontecer de cada membro da equipe possuir sua própria interpretação da imagem. Todavia, o principal problema é que para a recuperação de imagens o importante não é o que o indexador pensa que a imagem representa, mas sim o que o usuário associa à imagem. Essa associação depende do objetivo de usuário.

Por conseguinte, as duas principais desvantagens de se permitir consultas feitas apenas com o uso de descrições textuais são os esforços requeridos para a indexação e a ambigüidade destas descrições.

Esta abordagem é muito eficiente em um domínio no qual cada objeto possui uma única descrição. Um exemplo de aplicação seria um sistema para consultar figuras de animais, onde cada imagem tem uma única descrição e tanto o indexador quanto o usuário possuem a mesma percepção sobre o objeto [30].

4.2.5 Consulta por Características

Consulta Direta por Características da Imagem é o método de recuperação onde a consulta consiste de valores diretamente derivados da imagem. Uma característica é uma abstração de uma imagem a valores numéricos que um computador pode processar.

A principal vantagem desse método é que as abstrações podem ser derivadas das imagens automaticamente e objetivamente. A desvantagem é que a formulação de consultas é difícil para o usuário porque as condições têm que conter valores das características das imagens. Além disso, aplicações de sistemas de recuperação usando características das imagens são usualmente restritas a domínios específicos para reduzir a complexidade do modelo de extração de característica necessário.

Um exemplo de aplicação pode ser uma base de dados é o RUI (*Representation for Understanding Images*) [11], um sistema tutorial inteligente que oferece facilidades para estudantes de medicina estudarem grandes bancos de dados de imagens radiológicas. Os conceitos visuais são representados por características de forma, tamanho e localização de componentes anatômicos.

4.2.6 Métodos de Consulta em Bibliotecas Digitais

A maioria das Bibliotecas Digitais utiliza os métodos de consulta por browsing e por texto. As buscas por texto podem ser: *fulltext*, onde a consulta é comparada com todas as palavras no texto inteiro, sem distinguir a função das várias palavras ou *fielded searchig*, onde são identificados campos bibliográficos ou estruturais (tal como autor) e permitidas buscas por campos específicos ("autor" = "João"). Esses métodos são bastante poderosos, muito usados atualmente, freqüentemente, em combinação.

Entretanto como foi visto, quando da utilização desses métodos para indexação e buscas por imagens, muitos problemas se encontram, entre eles a dificuldade de descrever o conteúdo da imagem textualmente e a ambigüidade e subjetividade dos profissionais da descrição e indexação.

Dessa forma esse trabalho propõe a utilização das buscas por conteúdo nas Bibliotecas Digitais. Nossa proposta combina as consultas por imagem exemplo e por características. Essa combinação é feita da seguinte forma: o usuário entra com uma imagem exemplo e quer buscar imagens similares; para tanto,

o sistema extrai dessa imagem algumas de suas características visuais, as quais terão suas similaridades calculadas e comparadas junto às imagens (entende-se metadados das imagens) disponíveis na base de dados.

Essa combinação une as principais vantagens desses dois métodos: a utilização dos atributos visuais das imagens para a comparação entre elas e a extração automática e objetiva desses atributos. Sendo a consulta por imagem exemplo (na área médica, é comum que os profissionais possuam imagens para serem usadas como exemplo), elimina-se a desvantagem do método de consulta por características, onde a formulação das consultas impõe uma certa dificuldade para os usuários.

4.3 Buscas por Similaridade a Imagens Médicas

As consultas por similaridade são as principais ferramentas utilizadas em sistemas de recuperação de imagens baseada em conteúdo, sendo muito importantes em sistemas que manipulam imagens. É muito difícil acontecer de uma aplicação precisar comparar imagens iguais, o que seria feito pela comparação *pixel* a *pixel* entre elas. O que se deseja, na maior parte dos casos, é a pesquisa por imagens que sejam parecidas ou similares [3].

Assim, as consultas por similaridade são um dos mais importantes benefícios da aplicação da ciência da computação na área médica como ferramenta de apoio ao diagnóstico, sendo usualmente aplicadas na manipulação de imagens médicas tais como, raio-x, tomografia, ultra-som, ressonância magnética, mamografia, dentre outras [8].

Ao se trabalhar com um Sistema Gerenciador de Banco de Dados (SGBD) contendo, por exemplo, o cadastro de pacientes de um hospital, é comum procurar dados considerando algum critério de filtragem. Um exemplo simples de consulta seria "obter os resultados dos exames de sangue de todos os pacientes com dengue que foram atendidos após o início do último verão". No caso, o critério é composto pela especificação de uma doença [doença = "dengue"] e de um intervalo de tempo [data-atendimento $> 21/12/2001$]; a resposta fornecida pelo SGDB é composta pelos resultados dos exames de sangue em conformidade com as condições especificadas. Critérios como este são caracterizados por envolver igualdade, onde o interesse é por dados exatamente coincidentes e ordem, onde o interesse é por dados maiores ou menores que um valor fornecido. Os tipos de dados em questão são ditos convencionais e incluem basicamente valores numéricos, data/hora e cadeias de caracteres.

No entanto, critérios baseados em igualdade e ordem não são adequados aos chamados tipos de dados complexos, ou não convencionais, que são estruturalmente mais sofisticados. Exemplos na área médica incluem: imagens médicas em geral, cadeias de DNA, eletrocardiogramas, dentre outros. Para estes tipos não há sentido realizar consultas como "obter o cadastro dos pacientes com tumor no cérebro cuja tomografia seja igual à do paciente em estudo". Dificilmente as tomografias de dois tumores serão exatamente iguais, mesmo que os tumores tenham a mesma classificação. Portanto, o critério mais adequado para casos assim é o de semelhança. Assim, a consulta acima faria mais sentido da seguinte forma: "obter o cadastro dos pacientes com tumor no cérebro cuja tomografia seja bastante similar à do paciente em estudo".

Desse modo, fica clara a importância das buscas por similaridade em sistemas na área médica, justificando a importância dessas consultas também às imagens médicas das Bibliotecas Digitais, sobretudo através de sistemas especializados como os Provedores de Serviços (mais detalhes sobre os Provedores de Serviços serão apresentados no Capítulo 6). Com os metadados das imagens enriquecidos com as características visuais extraídas por algoritmos, um grande passo inicial será alcançado para a implementação desses sistemas.

Capítulo 5

Processamento de Imagens e Extração de Características

Um sistema genérico para processamento e análise de imagens é composto pelos seguintes módulos [7]:

- **Aquisição e digitalização de imagens**: consiste em transformar documentos em imagens digitais sob a forma de matrizes de valores chamados *pixels*.
- **Pré-processamento**: consiste em aprimorar a qualidade da imagem para as etapas subseqüentes. A imagem resultante desta etapa é uma imagem digitalizada de melhor qualidade que a original.
- **Segmentação**: consiste em identificar *pixels* com características semelhantes e agrupá-los em regiões relevantes da imagem. Segmentação significa agregar ou classificar regiões semelhantes da imagem.
- **Extração de Características**: consiste em extrair características das regiões resultantes da segmentação. Vale ressaltar que nesta etapa a entrada ainda é uma imagem, mas a saída é um conjunto de dados correspondentes àquela imagem.
- **Reconhecimento e Interpretação**: reconhecimento é o processo de atribuição de um rótulo a um objeto baseado em suas características traduzidas por seus descritores. A interpretação, por outro lado, consiste em atribuir significado a um conjunto de objetos reconhecidos.

Este trabalho teve o objetivo de estudar mais detalhadamente a segmentação e extração de características, uma vez que essas etapas são as responsáveis pela extração das características das imagens, que posteriormente poderão compor os metadados das imagens de forma automática e ainda permitir a recuperação dessas imagens em Sistemas de Recuperação de Imagens por Conteúdo.

5.1 Segmentação de Imagens

A segmentação é o primeiro passo na análise de imagens. A segmentação de imagens consiste na extração ou identificação de objetos de interesse contidos em uma imagem, onde o objeto é toda região com conteúdo semântico relevante para a aplicação desejada. Após a sementação, cada objeto é descrito através de suas propriedades geométricas e topológicas. Por exemplo, atribuitos como área, forma e textura podem ser extraídos dos objetos e posteriormente utilizados nos processos de análise [28].

Esta é, geralmente, uma das tarefas mais difíceis e pode determinar o sucesso ou fracasso da análise de imagens. Isto porque todos os cálculos da análise de imagens serão realizados sobre as regiões identificadas nesta etapa. A segmentação é complexa pois tenta traduzir para o computador um processo cognitivo extremamente sofisticado realizado através da visão humana.

A segmentação possibilita a extração de características relativas às formas contidas em uma imagem. Para isso, ela baseia-se nas características dos *pixels* de uma imagem para subdividí-la em partes e/ou objetos constituintes, conforme um critério de descontinuidade ou similaridade [38]. Quando se utiliza, como critério, a descontinuidade, a abordagem é particionar uma imagem baseando-se nas mudanças abruptas na função da imagem, como brilho ou profundidade. As principais áreas de interesse dentro dessa categoria são:

- **Detecção de pontos**: É a mais simples técnica de detecção. Um ponto terá uma mudança drástica do valor de cinza em relação aos seus vizinhos.
- **Detecção de linhas**: É o processo mais complicado, pois é necessário achar os pixels que são semelhantes e testá-los para verificar se são partes de uma linha comum.
- **Detecção de bordas**: É uma das técnicas básicas utilizadas pela visão humana no reconhecimento de objetos. É o processo de localização e realce dos *pixels* de borda, aumentado o contraste entre a borda e o fundo. Este processo verifica a variação dos valores de luminosidade de uma imagem.

Quando se utiliza a similaridade, as principais abordagens são:

- **Limiarização**: É uma técnica que consiste em converter uma imagem original (com mais de dois níveis de cinza) para uma imagem binária (apenas dois níveis : 0 - preto e 1 - branco).
- **Crescimento de Regiões**: É uma técnica que agrega *pixels* ou subregiões em regiões maiores. Uma abordagem simples é a agregação de *pixels* que começa com um conjunto *semente* e, a partir dele, cresce as regiões anexando a cada ponto semente aqueles *pixels* que possuam propriedades similares.

5.2 Extração de Características

Uma vez que uma imagem tenha sido segmentada em regiões, os agrupamentos resultantes de *pixels* segmentados são normalmente representados e descritos em um formato apropriado para o processamento posterior. Uma região pode ser representada em termos de suas características externas (bordas) ou em

termos de suas características internas (os *pixels* que compõem a região). A descrição da região depende da representação adotada. Por exemplo, uma borda pode ser descrita por características tais como comprimento ou número de concavidades. É importante ressaltar que as características escolhidas como descritores devem preferencialmente ser pouco afetadas por transformações como mudança de escala, rotação e translação [14].

Uma característica é uma função de uma ou mais medidas, calculadas de forma que quantifique alguma propriedade de um objeto. Logo, o processo de extração de características é o cálculo de valores que descrevem alguma propriedade dos objetos. A Figura 5.1 apresenta o processo de extração de características. Este processo produz o Vetor de Características de uma imagem.

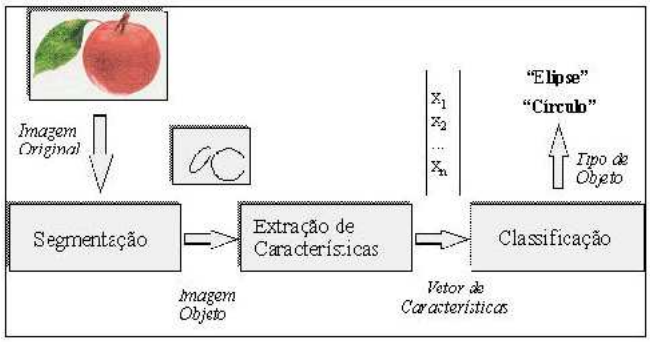

Figura 5.1: Processo de extração de características de uma imagem.

A extração de características é uma técnica comum para tratar a similaridade entre imagens. Uma imagem é descrita por um conjunto de muitas características, tais como, forma, textura e distribuição de cores. É através dessas características que as imagens são indexadas. Esse processo de extração é crucial para a armazenagem e recuperação das imagens baseada em seu conteúdo, permitindo sintetizar propriedades inerentes da imagem, que serão utilizadas no processo de indexação e recuperação das mesmas [33].

Abaixo se segue uma breve descrição das características de cor, textura e forma.

Cor

As cores presentes em uma imagem possuem um papel bastante significativo na indexação e recuperação da mesma. Existem diferentes representações de cores que incluem desde o tradicional RGB (*red, green, blue*), o mais simples modelo que mapeia diretamente as características físicas do dispositivo de exibição, até o HSI (*hue, saturation, intensity*) que reflete mais precisamente o modelo de cores para a percepção humana. Na realidade, todas as cores exibidas são criadas por combinações de quantidades apropriadas de vermelho, verde e azul. Um *pixel* de 24 bits em padrão RGB representa 224 ou aproximadamente 16.7 milhões de cores diferentes.

Muitas vezes, para aumentar a eficiência no processamento, as cores da imagem são re-quantizadas de forma a diminuir o número de cores possível e facilitar o tratamento das mesmas através de seu histograma [41]. O histograma de cores calcula e apresenta o número de *pixels* de uma imagem para cada

cor, ou seja, apresenta a distribuição das cores existentes na imagem digital em cada *pixel*.

Dois histogramas de cores podem ser comparados pelo somatório de diferenças absolutas ou quadráticas sobre o número de *pixels* de cada cor. Tal esquema é bastante simples e tolerante a pequenas alterações na imagem. Dessa forma, é natural que os histogramas de cores venham sendo estudados e implementados em sistemas de recuperação de imagens baseada em conteúdo. A popularidade da utilização de histograma de cores em sistemas de recuperação de imagens baseada em conteúdo deve-se, principalmente, a três fatores:

- É computacionalmente simples e barato calcular histogramas de cores.
- Pequenas alterações de movimentação na imagem pouco afetam os histogramas.
- Objetos distintos geralmente possuem histogramas diferentes.

Entretanto, não é possível separar ou reconhecer imagens utilizando somente o histograma de cores das mesmas, pois duas ou mais imagens bastante diferentes podem ter histogramas semelhantes. Ou seja, não há uma correspondência biunívoca entre a imagem e seu histograma de cores, levando ao surgimento do problema de ambigüidade. Tal fato é exemplificado na Figura 5.2. As quatro imagens (a), (b), (c), (d) possuem o mesmo histograma associado, apresentado em (e):

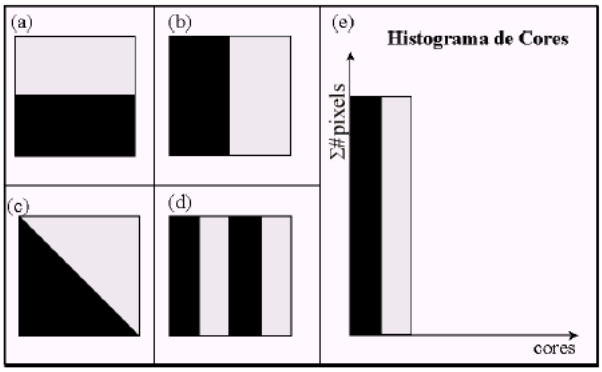

Figura 5.2: Mesmo histograma de cores (dois níveis de cinza) associado a quatro imagens distintas.

Devido ao caráter ambíguo do histograma de cores de uma imagem, outros métodos devem ser utilizados juntamente. Outro problema dos histogramas é que, como o número de cores das imagens geralmente é grande (mais de 100 níveis), indexar vetores dessa dimensão é problemático (um histograma para 100 cores distintas pode ser visto como um ponto 100-dimensional) [3].

Textura

Um elemento de textura é uma região de intensidade uniforme de formas simples que se repete dentro de um intervalo. Segundo Meyer [22] não existe uma definição simples e sem ambigüidades para textura, e a razão para isso é o conceito fortemente intuitivo de textura, impossível de ser totalmente descrito em uma definição textual.

O tratamento de textura difere do realizado sobre cores devido ao fato de que as texturas são definidas sobre regiões da imagem, e não sobre os pixels como as cores. A segmentação de uma imagem utilizando textura determina quais regiões da imagem possuem textura uniforme. Porém da mesma forma que o histograma de cores, há os mesmos problemas de ambigüidade e dimensionalidade, para a indexação de dados de textura.

Forma

A característica de forma é importante na representação do conteúdo de imagens médicas, permitindo, consequentemente, recuperação baseada em conteúdo.

A recuperação de imagens baseada em forma é um dos problemas mais difíceis de serem tratados por sistemas de recuperação de imagens por conteúdo. Isso porque é difícil segmentar os objetos de interesses presentes na imagem, limitando a recuperação por formas aos objetos mais bem discriminados presentes nas imagens.

A imagem é processada para buscar objetos de interesse. Após o objeto ser encontrado, sua borda precisa ser detectada utilizando algoritmos de Detecção de Bordas. O processo de detecção de bordas e formas fica mais difícil e comprometido em cenas complexas, nas quais há, além de ruídos, oclusão parcial de objetos e sombras sobre regiões das imagens.

5.3 Extração de Características de Imagens Médicas

Basicamente, a maioria dos sistemas que lidam com imagens usa características de distribuição de cor. As técnicas mais populares são as baseadas nos histogramas de cor ou brilho, devido à simplicidade de obtê-los e compará-los, pois as operações são executadas em tempo linear [33]. Entretanto, como muitas das imagens médicas não possuem cores, essas propriedades não são as mais importantes nessa área. Exceções se fazem quando são utilizadas fotografias, tais como na área da Dermatologia [23].

Para esses tipos de imagens, acredita-se que características baseadas em textura e forma das regiões obtidas da imagem podem discriminar e separar as imagens de um modo mais apurado [2]. Diversos estudos encontrados na literatura apresentam técnicas baseadas em ambas as abordagens, tais como [17, 39] para a textura e [35, 46] para a forma. Assim, características como texturas e formas ganham importância para a recuperação de imagens médicas e serão tratadas mais detalhadamente nas seções a seguir.

5.4 Principais Descritores de Forma

Há basicamente dois tipos de descritores de forma: baseados em bordas e baseados em regiões. Dentre os descritores de forma baseados em bordas estão o perímetro, a assinatura, etc. Dentre os descritores baseados em regiões estão a área, o centróide, os eixos principais, os descritores topológicos, os momentos, etc [28].

5.4.1 Perímetro

O perímetro de uma borda é um de seus descritores mais simples. O perímetro é a medida que descreve o comprimento do contorno de uma região. Em uma definição simples é o número de *pixels* do contorno de uma região R, sendo que um *pixel* está no contorno se possui algum vizinho que está fora da região R [38]. O perímetro é um descritor invariante quanto à translação e rotação, mas depende da mudança de escala das imagens.

5.4.2 Assinatura

Uma assinatura é uma representação unidimensional da borda de um objeto. As assinaturas geradas pela abordagem mais simples são invariantes quanto à translação, mas dependem de rotação e mudanças de escala.

5.4.3 Área

A área de uma região pode ser expressa como o número de *pixels* compreendendo a região. Embora a área possa ser usada como descritor, ela é geralmente aplicada a situações em que a escala dos objetos não varia.

5.4.4 Centróide

O centróide é o centro da massa de uma região, na forma de duas coordenadas, x e y. Seus valores são invariantes às mudanças de escala, rotação e translação.

5.4.5 Eixos Principais

Os eixos principais de uma região são os auto-vetores da matriz de covariância obtida usando-se os *pixels* da região como variáveis aleatórias. Os dois auto-vetores apontam na direção de espalhamento máximo da região, respeitando a condição de serem ortogonais. O grau de espalhamento é medido pelos auto-valores correspondentes. Portanto, o espalhamento e a direção principal de uma região podem ser descritos pelo maior auto-valor e seu auto-vetor correspondente. Esse tipo de descrição é invariante à rotação, mas depende de mudanças na escala se os auto-valores forem usados para medir o espaçamento.

5.4.6 Descritores Topológicos

As propriedades topológicas são úteis para descrições globais de regiões no plano da imagem. Uma possível descrição topológica é dada pelo número de buracos na região. Esta propriedade não será afetada por rotações ou escala.

Outra propriedade topológica útil para descrição de regiões é o número de componentes conexos. Um componente conexo é um subconjunto de tamanho máximo tal que quaisquer dois pontos nesse subconjunto possam ser unidos por uma curva conexa que também pertença completamente ao subconjunto.

O número de buracos H e de componentes conexos C em uma figura podem ser usados na definição do número de Euler E:

$$E = C - H$$

O número de Euler é também uma propriedade topológica.

5.4.7 Momentos

A forma dos segmentos da borda pode ser descrita quantitativamente através de *momentos*. Um momento m é o número total de pontos na região, ou seja, equivale à área da região.

Propriedades de invariância quanto às transformações de escala, translação e rotação (*RST-invariant*)podem ser derivadas uilizando funções de momentos, as quais geram um conjunto de sete momentos invariantes.

5.5 Principais Abordagens de Textura

Uma importante característica para a descrição de regiões é a quantificação de seu conteúdo de textura. Embora não exista nenhuma definição formal de textura, esse descritor intuitivamente fornece medidas de propriedades como suavidade, rugosidade e regularidade. As três abordagens mais utilizadas na descrição de texturas são: estatística, estrutural e espectral [14].

5.5.1 Abordagem Estatística

Nessa abordagem, a textura é definida por um conjunto de medidas locais extraídas do padrão. Medidas estatísticas comuns incluem entropia, correlação, contraste e variância.

5.5.2 Abordagem Estrutural

Essa abordagem utiliza a idéia de que texturas são compostas de primitivas dispostas de forma quase regular e repetitiva, de acordo com regras bem definidas. Como exemplo, pode-se citar a descrição da textura baseada em linhas paralelas regularmente espaçadas.

5.5.3 Abordagem Espectral

Essa abordagem baseia-se em propriedades do espectro de Fourier, sendo usadas basicamente na detecção de periodicidade global em uma imagem através da identificação de picos de alta energia no espectro.

Capítulo 6

Sistemas de Buscas aos Metadados das Bibliotecas Digitais

Como visto no Capítulo 4, esse trabalho propôs um mecanismo de preenchimento dos metadados das imagens das Bibliotecas Digitais com conteúdo visual dessas imagens, extraído por algoritmos. Essas características serão usadas para descrever e indexar as imagens, possibilitando a comparação entre elas em Sistemas de Recuperação de Imagens por Conteúdo. Sendo assim, sistemas que promovem buscas aos metadados das Bibliotecas Digitais, tais como os Provedores de Serviços, conseguirão dar suporte às Buscas por Similaridade, uma importante ferramenta em sistemas que tratam de imagens, sobretudo, de imagens médicas.

Provedores de Serviços são sistemas participantes da Iniciativa de Arquivos Abertos, chamada OAI. A OAI é uma iniciativa com o intuito de prover meios para a interoperabilidade entre Bibliotecas Digitais, para que elas possam compartilhar seus conteúdos. Para tanto a Iniciativa desenvolveu um protocolo chamado OAI-PMH, responsável por colher os metadados dos repositórios (também chamados Provedores de Dados), como, por exemplo, as Bibliotecas Digitais, oferecendo uma técnica simples para que esses repositórios tornem seus metadados disponíveis aos Provedores de Serviços. Para tanto, um acordo entre Provedores de Dados e Provedores de Serviços deve ser estabelecido com relação ao formato de metadados utilizados por ambos.

Dessa forma, uma vez que um novo formato de metadados é utilizado pelos participantes da Iniciativa (pois como proposto no trabalho, novos elementos descritivos serão inseridos nos metadados das imagens), tarefas são necessárias para propor junto à OAI um novo esquema com novos elementos de metadados.

6.1 Open Archives Initiative (OAI)

A *Open Archives* é uma iniciativa para habilitar o acesso aos materiais eletrônicos da *Web*, através da interoperabilidade entre repositórios, para compartilhamento, publicação e arquivamento de metadados [24], diminuindo a barreira da interoperabilidade entre repositórios razoavelmente heterogêneos e facilitando a disseminação eficiente dos conteúdos digitais.

A OAI é formada por um comitê-diretor, composto por conceituadas universidades e outras entidades com interesse na disponibilização digital da informação, e por um comitê técnico presidido por Carl Lagose e Herbert Van de Sompel. A Iniciativa é patrocinada pela D*igital Library Federation, Coalition for Networked Information* e *Natural Science Foundation* [40].

Os participantes da iniciativa são divididos em Provedores de Dados e Provedores de Serviços. Os Provedores de Dados mantêm repositórios de documentos digitais que dão suporte ao protocolo OAI-PMH como forma de expor os metadados de seus documentos. Já os Provedores de Serviços oferecem buscas a esses metadados e outros serviços que visam agregar valor à Iniciativa. No âmbito da OAI, Bibliotecas Digitais são Provedores de Dados que exportam seus metadados aos Provedores de Serviços.

Os participantes da OAI registram suas instituições como Provedores de Dados ou Serviços no site da OAI, onde também podem ser encontradas a listas das instituições já registradas.

6.2 OAI Protocol for Metadata Harvesting (OAI-PMH)

O OAI-PMH (Protocolo OAI para Colheita de Metadados) foi criado com o intuito de oferecer simplicidade e eficiência na tarefa de unificar as consultas às bases de dados científicas/acadêmicas. Com os recursos oferecidos pelo protocolo é possível melhorar significativamente a precisão das consultas eletrônicas e reduzir o tempo de procura, graças ao compartilhamento de informações (metadados) entre os participantes da Iniciativa.

O OAI-PMH define um mecanismo para colher os metadados dos repositórios, oferecendo uma técnica simples para que os fornecedores de dados (Provedores de Dados) tornem seus metadados disponíveis a serviços baseados nos padrões abertos HTTP (*Hypertext Transport Protocol*) [16] e XML (*Extensible Markup Language*) [44].

Os metadados que são colhidos podem estar em qualquer formato que se encontre de acordo com uma comunidade (ou com um conjunto de provedores de dados e serviços), embora o *Unqualified* Dublin Core (metadados Dublin Core que não utilizam qualificadores, somente os quinze elementos do conjunto DCMES) deva ser o formato utilizado para fornecer um nível básico de interoperabilidade.

Através do OAI-PMH, metadados de quaisquer fontes podem ser colhidos e agrupados em uma única base de dados, e os serviços podem ser fornecidos baseados em um "coletor central", ou nos "dados agregados".

6.3 Definições Chaves da OAI

- **Arquivo** : O termo "arquivo" do nome "Iniciativa de Arquivos Abertos" é usado no sentido de repositório para armazenamento de informações. Um repositório é um servidor acessível por uma rede de computadores capaz de processar as requisições do protocolo OAI-PMH corretamente.

- **Protocolo**: Um protocolo é um conjunto de regras definindo a comunicação entre sistemas. FTP (*File Transfer Protocol*) e HTTP (*Hypertext Transport Protocol*) são exemplos de protocolos usados para comunicação entre sistemas através da Internet.

- ***Harvesting* (Colheita)**: No contexto da OAI, o termo "colheita" refere-se especificamente ao agrupamento de metadados, originados de vários repositórios distribuídos, em um único local.

- **Interoperabilidade**: Interoperabilidade é a habilidade dos sistemas, organizações e serviços de trabalharem juntos através de técnicas diversas ou comuns. Tecnicamente ela é sustentada por padrões abertos para comunicação entre sistemas e para descrição de recursos e coleções, entre outros. Em termos de metadados, interoperabilidade significa que eles podem ser compartilhados entre organizações, aumentando o escopo das coleções e a possibilidade de implementação de sistemas de buscas entre essas coleções. Nesses casos, a interoperabilidade é considerada principalmente no contexto de descoberta e acesso aos recursos.

- **Registro**: É um metadado escrito em um determinado formato.

- **XML**: A XML (*eXtensible Markup Language*) é uma linguagem para criação de outras linguagens, definindo um meio para descrição de dados. Um documento XML pode ser validado junto a um DTD (*Document Type Definition*) ou a um Esquema que descreva sua estrutura (os elementos da linguagem criada, seus relacionamentos, atributos, etc). Um DTD ou um Esquema XML é uma especificação formal da estrutura do documento XML relacionado [4].

- **Recurso**: Um recurso é algo que possui uma identificação. Por exemplo, um documento eletrônico, uma imagem, ou uma coleção de recursos. Nem todos os recursos são recuperáveis por uma rede de computadores, dessa forma, pessoas, corporações e livros de uma biblioteca também podem ser considerados recursos.

- **Namespaces XML**: Um namespace XML é uma coleção de nomes, identificados por uma referência URI, utilizada em documentos XML como tipos de elementos e nomes de atributos [19].

- **Esquemas XML**: Esquemas XML definem os elementos, os relacionamentos entre eles, o tipo de conteúdo dos elementos, entre outras coisas, que compõem os documentos XML [36]. Eles fornecem um significado à estrutura definida, ao conteúdo e à semântica dos documentos XML.

- **Esquemas Recipientes (*Containers Schemas*)**: Esquemas Recipientes são os locais onde os documentos XML obtidos como respostas às requisições do protocolo OAI-PMH podem ser validados. As diretrizes de implementação da OAI listam os recipientes opcionais existentes e fornecem links aos mesmos.

6.4 Provedores de Dados

No âmbito da OAI, Provedores de Dados são bases de dados (ou repositórios) que exportam os metadados dos seus documentos digitais. Como o que é exportado é o metadado do registro, seu conteúdo (por exemplo, texto completo, imagem ou vídeo) não é necessariamente exposto, a menos que o repositório inclua o endereço digital (*link*) do documento em um dos campos dos metadados. A ligação entre o metadado e seu conteúdo não é definida pelo protocolo. Assim, é importante ressaltar que o protocolo OAI-PMH não fornece uma busca aos dados, ele simplesmente torna possível trazê-los juntos para algum lugar. Para fornecer serviços adicionais, a técnica de colheita deve ser combinada com outros mecanismos.

6.5 Provedores de Serviços

Um Provedor de Serviços faz as requisições do protocolo OAI-PMH aos Provedores de Dados, para colher os metadados e utilizá-los como base para a construção de serviços.

São os Provedores de Serviços que coletam, organizam e disponibilizam os metadados dos Provedores de Dados. Ao acessar um Provedor de Serviços e realizar uma pesquisa, a cada dia se tem respostas mais precisas, dada à crescente adesão à Iniciativa por parte das comunidades científicas.

Como exemplo de um provedor de serviços tem-se o OAIster [25], implementado pela University of Michigan Digital Library Production Services, de onde é possível acessar mais de 160 provedores de dados registrados na OAI de um único ponto. Atualmente a interface para os usuários buscarem os metadados no OAIster pode ser vista na Figura 6.1.

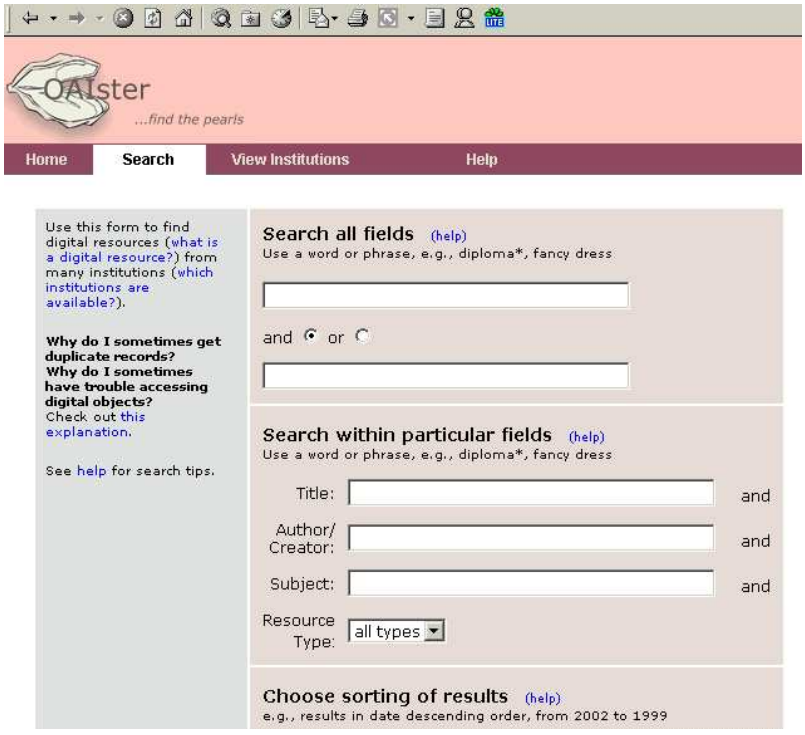

Figura 6.1: Interface de Busca no OAIster.

Como pode ser visto na Figura 6.1, as buscas podem ser feitas através de frases ou palavras que serão pesquisadas em todos os campos dos metadados, ou através dos campos específicos: Título, Autor, Assunto e Tipo (texto, imagem, áudio ou vídeo). No caso de haver informações sobre o conteúdo das imagens presentes nos metadados colhidos pelo protocolo OAI-PMH, essa interface poderia ser alterada, permitindo Buscas por Conteúdo, onde o usuário pode entrar com uma imagem exemplo para obter imagens similares.

6.6 Descrição de Imagens por Conteúdo nos Provedores de Dados

Para que os Provedores de Dados criem metadados que poderão permitir recuperação de imagens por conteúdo nos Provedores de Serviços, algumas questões são importantes.

Inicialmente é necessária a definição de um Modelo de Representação do Conteúdo das Imagens. Essa definição consiste em determinar características relevantes que possam ser extraídas das imagens para melhor representar o conteúdo das mesmas.

Dessa forma, Bibliotecas Digitais participantes da OAI que queiram indexar automaticamente suas imagens, habilitar buscas por conteúdo e ainda compartilhar seus metadados, devem entrar em comum acordo quanto as características visuais inseridas nos metadados (posteriormente declaradas em um esquema XML). Uma vez que tais Bibliotecas compartilham do mesmo modelo de representação do conteúdo das imagens, a interoperabilidade está garantida. Em outras palavras, se cada Biblioteca Digital utilizar diferentes características para representar o conteúdo de suas imagens, seus metadados possuírão diferentes elementos (seguirão diferentes esquemas XML) e a interoperabilidade será afetada, prejudicando também os sistemas de buscas a esses metadados.

Uma vez definidas essas características, deve-se pensar como elas serão inseridas nos documentos XML. Isso inclui, quais elementos XML deverão ser criados e como eles serão inseridos dentro da estrutura do documento, junto aos demais elementos Dublin Core. Pensando nisso, a OAI deve criar um novo esquema XML, onde tais elementos serão declarados, fazer a validação do mesmo junto ao protocolo OAI-PMH e deixá-lo diponível para que o protocolo possa utilizá-lo para validar os metadados colhidos dos Provedores de Dados. Mais informação sobre a validação desse esquema estarão disponíveis na seção 6.8.

6.7 Recuperação de Imagens por Conteúdo nos Provedores de Serviços

De acordo com o método proposto, os metadados colhidos dos Provedores de Dados através do protocolo OAI-PMH, que chegam aos Provedores de Serviços, possuírão atributos visuais das imagens, tornando possível buscas baseadas em conteúdo a essas imagens.

Entretanto, para que esse tipo de busca seja possível, é necessário que os Provedores de Serviços implementem módulos de recuperação de imagens por conteúdo, os quais deverão ser capazes de calcular medidas de similaridade entre a imagem de busca fornecida pelo usuário e as imagens representadas nos metadados disponíveis na base de dados desse provedor, cujos conteúdos visuais estão representados na forma de vetores de características.

A Figura 6.2 ilustra essas tarefas.

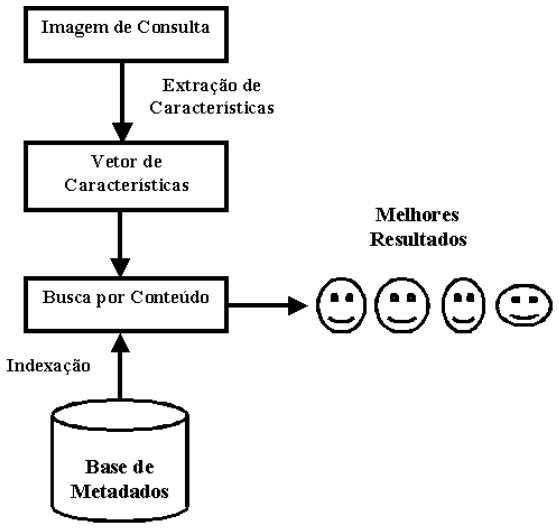

Figura 6.2: Busca por Conteúdo.

Por essa figura, pode-se ver que inicialmente o usuário deve entrar com uma imagem de consulta para encontrar imagens similares. Tal imagem é processada por algoritmos apropriados, e um vetor de características é gerado. Tal vetor será comparado com os vetores presentes nos metadados disponíveis na base de dados, para que o módulo de recuperação por conteúdo faça os cálculos de similaridade entre tais vetores, e devolva ao usuário um conjunto de imagens resultantes similares à imagem de consulta.

Apesar da Figura 6.2 não ilustrar a busca textual, ela ainda pode ser utilizada para buscar as imagens, uma vez que os metadados possuem elementos textuais. O Photobook [29], o MARS [34] e o EPIC [21] são exemplos de SRICs que utilizam tanto o conteúdo das imagens quanto anotações textuais para recuperarem suas imagens.

6.8 Esquemas XML e o Suporte para Múltiplos Formatos de Metadados no Protocolo OAI-PMH

O protocolo OAI-PMH utiliza Esquemas XML (*XML Schemas*) para definir o formato dos metadados dos registros a serem colhidos. O esquema XML desenvolvido pela OAI para validar os metadados a serem colhidos é chamado oai-dc. O oai-dc é um formato de metadados simples, cujos elementos são baseados no Dublin Core *Unqualified*. Este esquema é usado como o formato de metadados mínimo para interoperabilidade exigido pelo protocolo OAI-PMH. Isto quer dizer que, a menos que determinado Provedor de Dados defina seu próprio formato de metadados através dos Esquemas XML e valide-o junto à OAI, os metadados dos seus registros deverão estar de acordo com o formato definido no oai-dc.

6.8.1 Usando Outros Esquemas de Metadados

O oai-dc é um formato simples que fornece um nível básico de interoperabilidade. Porém, há inúmeras razões pelas quais ele pode não ser adequado para determinados repositórios, serviços ou comunidades:

- **Os 15 elementos DCMES podem não incluir os elementos necessários para um determinado domínio.** Neste caso pode-se criar um novo esquema incorporando os elementos adicionais necessários entre aqueles já utilizados no DCMES.

- **Os elementos do oai-dc podem não ser suficientemente precisos para os registros de metadados de um determinado domínio**, sendo o DCMES um esquema de codificação de metadados "unqualified" (sem qualificadores). Neste caso pode-se obter uma maior precisão adicionando "encoding schemes" (vocabulários controlados, por exemplo, o código ISO3166 para representação de nomes de países) aos elementos DCMES existentes.

- **O Dublin Core pode não ser o formato de metadados necessário para uma determinada comunidade.** Em uma comunidade pode-se desejar trocar metadados em outro formato como, por exemplo, em IMS/IEEE LOM para metadados *eLearning*.

6.8.2 Adicionando Novos Elementos quando o oai-dc não é Suficiente

Criar um novo esquema estendendo o oai-dc, através da adição de novos elementos, envolve as seguintes tarefas:

1. Criar um nome para o novo esquema.
2. Criar namespaces.
3. Criar o esquema para os novos elementos.
4. Criar um esquema recipiente (*container schema*).
5. Validar o esquema/registros.
6. Informar aos verbos do OAI-PMH sobre o novo formato.
7. Testar se o esquema funciona e se é válido.

Quando todas essas condições são satisfeitas, tem-se um novo formato. Mais informações sobre cada uma dessas tarefas estão disponíveis no Anexo I.

Capítulo 7

Ilustração do Método Proposto

Para a ilustração do método proposto, inicialmente é necessária a escolha de um conjunto de descritores que formarão o vetor de características da imagem de forma a representar seu conteúdo e o processamento de tais imagens para a extração automática de tais descritores. Em seguida, deve-se pensar na nova estrutura que os metadados (documentos XML) irão possuir no momento em que esse vetor será inserido nos mesmos. Finalmente, deve-se desenvolver um aplicativo que deverá ser capaz de ler o vetor de características da imagem e inserí-lo automatimente nos metadados das imagens. Esse aplicativo será desenvolvido na linguagem XSLT, cujos documentos são chamados de *folhas de estilo*.

7.1 Esolha dos Descritores de Conteúdo das Imagens Médicas

As técnicas de extração de características são geralmente determinadas pelos tipos de variações presentes nas imagens. No caso de imagens médicas, esses tipos de variações podem depender da modalidade (raio-x, tomografias, etc), da área considerada (coração, crânio, tórax, etc), havendo a necessidade da definição de estratégias específicas para o processo de extração de características dessas imagens.

Pode-se adiantar que não existe uma "melhor" característica para prover uma representação precisa em um determinado conjunto de imagens, mesmo porque uma única característica geralmente não é suficiente para garantir a unicidade de uma imagem. Nesse caso, uma combinação de características pode prover uma representação que proporcione um resultado mais adequado na recuperação das imagens.

Portanto, na prática, a Recuperação de Imagens por Conteúdo combina várias características para melhorar a eficiência da recuperação. Essas características podem trabalhar de forma integrada, reduzindo o conjunto das imagens candidatas como respostas às consultas e propiciando, dessa maneira, uma melhor discriminação entre as imagens.

Uma vez que o objetivo principal desse trabalho é mostrar uma técnica para automatizar a criação dos metadados das imagens com características visuais extraídas por algoritmos, não foi parte do escopo definir estratégias específicas para o processo de extração de características de imagens médicas. Logo,

não foram realizados estudos profundos quanto aos descritores de características que melhor representam o conteúdo das imagens médicas utilizadas como exemplo. Entretanto, houve a preocupação de exemplificar o método com características bastante usadas em sistemas de recuperação de imagens médicas, tais como forma e textura.

Dessa maneira, para exemplificar o método de criação automática dos metadados das imagens proposto no trabalho, as características de forma e textura foram escolhidas, levando à realização de um estudo sobre os principais descritores dessas características, para definir quais iriam compor os metadados.

Uma questão importante para a escolha dos descritores de forma e textura para representar o conteúdo das imagens, foi com relação aos Sistemas de Recuperação de Imagens por Conteúdo a serem implementados pelos Provedores de Serviços. Os SRICs geralmente impõem a padronização dos equipamentos utilizados para a digitalização de imagens (equipamentos de diferentes padrões implicam diferenças em algumas medidas das imagens, tais como resolução, mudanças na coloração e qualidade gráfica). Se essas restrições não forem seguidas, as diferenças podem dificultar a comparação entre imagens e, logo, afetar a precisão das consultas.

Entretanto, no caso do método proposto junto aos participantes da iniciativa OAI, não seria aceitável que a iniciativa impusesse aos seus participantes esses tipos de restrições. Por exemplo, seria impossível impor que todas as Bibliotecas Digitais utilizassem imagens digitalizadas por um mesmo tipo de equipamento. Logo, é interessante utilizar descritores de características que sejam pouco afetados por transformações como mudanças de escala, rotação e translação.

Dessa maneira, de acordo com os estudos sobre representação do conteúdo de imagens médicas e de acordo também com as necessidades do método proposto com relação à recuperação por conteúdo nos Provedores de Serviços, foram escolhidos os seguintes descritores de forma e textura:

- Centróide
- Número de Euler
- Momentos
- Entropia
- Correlação

A ferramenta utilizada para o processamento das imagens a fim de extrair esses descritores foi a CVIPtools 3.9 [5], desenvolvida pelo *Computer Vision and Image Processing Laboratory* da *Southern Illinois University*. Através dessa ferramenta, inicialmente a imagem é segmentada para então ser possível o cálculo desses descritores.

7.2 Resultados Obtidos do Processamento das Imagens através da Ferramenta CVIPtools

Como exemplo, foram processadas três imagens médicas de tipos distintos: Raio-x, Ressonância Magnética e Tomografia. Para cada uma das imagens foi gerado um vetor de características cuja ordem dos elementos

é a seguinte:

1 - Centróide (*row*)

2 - Centróide (*column*)

3 - Número de Euler

4 - RST1

5 - RST2

6 - RST3

7 - RST4

8 - RST5

9 - RST6

10 - RST7

11 - Correlação (*average*)

12 - Correlação (*range*)

13 - Entropia (*average*)

14 - Entropia (*range*)

7.2.1 Imagem de Raio-X

Figura 7.1: Imagem de Raio-X

Vetor de características gerado:

489 301 -111 0.185917 0.002431 0.001080 0.000011 0.000000 0.000000 0.000000 0.955407 0.020630 4.765155 0.222435

7.2.2 Imagem de Ressonância Magnética

Vetor de características gerado:

233 200 260 0.365081 0.003821 0.000595 0.000402 0.000000 -0.000001 0.000000 0.759228 0.061941 5.803944 0.024866

Figura 7.2: Imagem de Ressonância Magnética

7.2.3 Imagem de Tomografia

Figura 7.3: Imagem de Tomografia

Vetor de características gerado:

289 258 -110 0.167794 0.000721 0.000031 0.000001 0.000000 0.000000 -0.000000 0.970453 0.014869 6.530753 0.258148

7.3 Desenvolvimento da Folha de Estilo

Os metadados das Bibliotecas Digitais são geralmente armazenados em documentos XML. Sendo assim, informações são adicionadas a esses metadados através da criação de novos elementos XML.

Para que os documentos XML sejam processados a fim de receberem novos elementos descritivos, uma linguagem que pode ser utilizada é a XSLT (*eXtensible Stylesheet Language for Transformations*) [45]. Documentos desenvolvidos em linguagem XSLT (chamados folhas de estilo) podem conter regras de processamento responsáveis por modificar documentos XML. No caso desse trabalho é necessário que

uma folha de estilo leia um vetor de característica gerado por um algoritmo e insira-o no documento XML correspondente ao metadado da imagem processada.

Entretanto, antes do desenvolvimento da folha de estilo, é necessário estabelecer qual será a nova estrutura (os novos elementos descritores) do documento XML resultante.

Para cada um dos descritores listados na seção 7.1 serão criados os seguintes elementos e atributos XML:

```
<centroid>
 <row></row>
 <column></column>
</centroid>

<eulernumber> </eulernumber>

<moment>
 <rst1></rst1>
 <rst2></rst2>
 <rst3></rst3>
 <rst4></rst4>
 <rst5></rst5>
 <rst6></rst6>
 <rst7></rst7>
</moment>

<correlation>
 <average></average>
 <range></range>
</correlation>

<entropy>
 <average></average>
 <range></range>
</entropy>
```

7.4 Processamento da Folha de Estilo

Para o processamento da folha de estilo desenvolvida, foi utilizado o processador Saxon [32]. O pacote Saxon é uma coleção de ferramentas para processamento de documentos XML.

Para as três imagens mostradas na seção 7.2, os metadados processados pela folha de estilo ficaram da seguinte forma:

Imagem de Raio-X

```
<?xml version="1.0" encoding="UTF-8"?>
<metadata>
 <dc:creator>Carlos, Antonio</dc:creator>
 <dc:title>X-Ray Image</dc:title>
 <dc:date>2004-03-02</dc:date>
 <dc:description>X-Ray Image of Hospital de Clínicas
 (UFPR)</dc:description>
 <dc:format>JPG</dc:format>
 <centroid>
 <row>489</row>
 <column>301</column>
 </centroid>
 <eulernumber>-111</eulernumber>
 <moment>
 <rst1>0.185917</rst1>
 <rst2>0.002431</rst2>
 <rst3>0.001080</rst3>
 <rst4>0.000011</rst4>
 <rst5>0.000000</rst5>
 <rst6>0.000000</rst6>
 <rst7>0.000000</rst7>
 </moment>
 <correlation>
 <average>0.955407</average>
 <range>0.020630</range>
 </correlation>
 <entropy>
 <average>4.765155</average>
 <range>0.222435</range>
 </entropy>
</metadata>
```

Imagem de Ressonância Magnética

```
<?xml version="1.0" encoding="UTF-8"?>
<metadata>
 <dc:creator>Carlos, Joao</dc:creator>
 <dc:title>Magnetic Resonance Image</dc:title>
 <dc:date>2004-03-02</dc:date>
 <dc:description>Magnetic Resonance Image of Hospital de Clínicas
```

```
 (UFPR)</dc:description>
 <dc:format> JPG</dc:format>
 <centroid>
 <row>233</row>
 <column>200</column>
 </centroid>
 <eulernumber>260</eulernumber>
 <moment>
 <rst1>0.365081</rst1>
 <rst2>0.003821</rst2>
 <rst3>0.000595</rst3>
 <rst4>0.000402</rst4>
 <rst5>0.000000</rst5>
 <rst6>-0.000001</rst6>
 <rst7>0.000000</rst7>
 </moment>
 <correlation>
 <average>0.759228</average>
 <range>0.061941</range>
 </correlation>
 <entropy>
 <average>5.803944</average>
 <range>0.024866</range>
 </entropy>
</metadata>
```

Imagem de Tomografia

```
<?xml version="1.0" encoding="UTF-8"?>
<metadata>
 <dc:creator>Carlos, Jose</dc:creator>
 <dc:title>Tomography Image</dc:title>
 <dc:date>2004-03-02</dc:date>
 <dc:description>Tomography Image of Hospital de Clínicas (UFPR)
 </dc:description>
 <dc:format>JPG</dc:format>
 <centroid>
 <row>289</row>
 <column>258</column>
 </centroid>
```

```
 <eulernumber>-110</eulernumber>
 <moment>
 <rst1>0.167794</rst1>
 <rst2>0.000721</rst2>
 <rst3>0.000031</rst3>
 <rst4>0.000001</rst4>
 <rst5>0.000000</rst5>
 <rst6>0.000000</rst6>
 <rst7>-0.000000</rst7>
 </moment>
 <correlation>
 <average>0.970453</average>
 <range>0.014869</range>
 </correlation>
 <entropy>
 <average>6.530753</average>
 <range>0.258148</range>
 </entropy>
</metadata>
```

Note que os novos elementos criados para a descrição do conteúdo da imagem são inseridos dentro do elemento *metadata*, uma vez que eles também são vistos como metadados das imagens.

Os elementos $dc:creator$, $dc:title$, $dc:date$, $dc:description$ e $dc:format$ do Padrão Dublin Core, são criados manualmente pelo profissional responsável por essa coleção da Biblioteca Digital.

Capítulo 8

Conclusão

Esse trabalho apresenta um amplo estudo sobre a utilização de metadados para a descrição de recursos. São apresentados os tipos de metadados existentes, sobretudo, os metadados descritivos, os quais são responsáveis por descrever o conteúdo dos recursos e prover formas de identificação e recuperação dos mesmos.

A aplicação dos metadados geralmente é controlada pelo uso de esquemas e especificações. O trabalho mostra a importância quanto à utilização de metadados que seguem determinada especificação, para que eles se tornem interoperáveis com outras instituições, facilitando a utilização e a recuperação dos mesmos. Entre as especificações existentes, foi estudado o Padrão Dublin Core, que é um padrão simples e muito utilizado para a descrição de recursos na *Web*. Embora o Dublin Core seja eficiente, quando da sua utilização para a descrição de determinados tipos de recursos, seus elementos nem sempre conseguem descrever todo o conteúdo do recurso, como ocorre, por exemplo, com imagens.

As imagens são os tipos de recursos discutidos no trabalho. Como foi mostrada, a criação de metadados descritivos de imagens é um processo delicado, que exige grande esforço manual de uma equipe especializada, custos e tempo. Ainda que a criação dos metadados e a indexação das imagens sejam feitas por pessoas qualificadas, essa geralmente não é a melhor solução, pois a descrição do conteúdo de uma imagem, quando feita somente por pessoas através de textos, não consegue alcançar um resultado completo.

Como exemplo de aplicação que utiliza metadados para a descrição de seus recursos, o trabalho estudou as Bibliotecas Digitais. As Bibliotecas Digitais são organizações de grande importância atualmente, fornecendo uma variedade de serviços, de uma forma eficiente e prática aos seus usuários. Uma questão muito importante com relação às Bibliotecas Digitais é a disponibilização de suas informações. Nessa área, um grande avanço tem sido alcançado graças à criação da iniciativa *Open Archives* (OAI). Essa iniciativa, tem o intuito prover o compartilhamento de informações entre repositórios de dados. Esse trabalho mostrou que, através dos recursos da OAI, é possível melhorar significativamente a precisão das

consultas e reduzir o tempo de procura aos dados.

Além disso, foram apresentados os participantes da OAI - Provedores de Dados e Provedores de Serviços -, onde ambos utilizam o protocolo OAI-PMH para expor os metadados (no caso dos Provedores de Dados) e para colher os metadados (no caso dos Provedores de Serviços). Os Provedores de Serviços foram vistos como sistemas altamente eficientes na tarefa de prover aos usuários buscas aos metadados das Bibliotecas Digitais, uma vez que eles conseguem unificar metadados de várias fontes e fornecê-los de forma única e transparente aos usuários.

Bibliotecas Digitais também são, no entanto, ambientes propícios às atividades derivadas da automatização tecnológica e manipulação digital dos documentos. Uma das linhas de pesquisa que mais têm sido estudadas diz respeito às possibilidades de indexação automática dos documentos ao entrarem em suas bases de dados, a partir de diversas estratégias que permitirão o oferecimento de novas formas de buscas semânticas sobre os documentos, somando novas possibilidades aos processos tradicionais de recuperação de informações. Com relação à indexação de imagens, foi visto que a criação e manutenção de índices para grandes coleções de imagens envolvem custos e tempo. A indexação de imagens por textos devolve somente uma resposta: sucesso ou fracasso. Se o usuário não especificar as palavras-chaves de uma forma correta, as imagens desejadas podem nunca ser recuperadas.

Dessa forma, a recuperação da informação pode se tornar mais eficiente se o computador for capaz de extrair automaticamente características visuais. Nesse caso, o computador analisa cada imagem e extrai informações tais como cor, textura e forma. Tais informações descrevem objetivamente as imagens e são usadas para criar índices automaticamente e para comparar as imagens durante o processo de recuperação baseada em conteúdo. Esse método garante uma abstração objetiva da imagem, pois não envolve a subjetividade humana. Logo, se os metadados das imagens contém características visuais é possível indexá-los automaticamente.

Portanto, a linha mestra desse trabalho foi a proposta de uma técnica para a criação automática dos metadados das imagens das Bibliotecas Digitais, e também de indexação automática das imagens ao entrarem para as bases de dados das Bibliotecas Digitais. Essa automatização se dá através da utilização de algoritmos de processamento de imagens, capazes de extraírem características que descrevem o conteúdo visual de uma imagem. Esses algoritmos são utilizados por Sistemas de Recuperação de Imagens por Conteúdo, uma vez que esses sistemas utilizam tais características para recuperar as imagens.

O trabalho estudou o funcionamento dos Sistemas de Recuperação de Imagens por Conteúdo, discutindo sobre as maneiras pelas quais as imagens podem ser indexadas, as dificuldades apresentadas, os métodos de consultas mais utilizados e as vantagens e desvantagens de cada um. Através dos resultados desses estudos, foi possível pensar em um método que una três técnicas de consulta a imagens: consulta textual, consulta por imagem exemplo e consulta por características. Para tanto, os metadados deveriam possuir informações textuais e visuais, e o usuário poderia fazer sua consulta através de textos e de imagens exemplos. No caso de se usar imagens exemplo, o sistema ficaria responsável por extrair características dessa imagem, e compará-las com as características presentes nos metadados disponíveis para buscas, para então encontrar um conjunto resposta de imagens que satisfaça o objetivo do usuário.

Com a comprovação das vantagens que essa junção poderia trazer, foi proposta a idéia de criar

metadados de imagens tanto com informações textuais, quanto com informações visuais. Esses metadados, então, permitiriam que as Bibliotecas Digitais e os sistemas que promovem buscas aos seus metadados, como os Provedores de Serviços, se tornassem sistemas capazes de prover Buscas por Similaridade. As Buscas por Similaridade foram vistas como uma importante ferramenta aos sistemas que lidam com imagens, como foi mostrado para o domínio de imagens médicas.

Para que fosse entendido um pouco mais sobre extração automática de características por algoritmos de processamento de imagens, o trabalho apresentou as etapas do processamento de imagens que utilizam tais algoritmos e as principais características extraídas. Entre as etapas do processamento estudadas estão a segmentação e extração de características. Entre as principais características estudadas estão cor, forma e textura.

Com as idéias do trabalho bem esclarecidas, foi proposto o desenvolvimento de um aplicativo em linguagem XSLT capaz de ler os chamados "vetores de características" gerados automaticamente por algoritmos e inserí-los nos documentos XML que armazenam os metadados das imagens das Bibliotecas Digitais.

Através dos resultados obtidos, foi mostrado que o método proposto é capaz de automatizar a criação dos metadados das imagens de forma rápida, eficiente e sem ambigüidades, e ainda prover indexação automática desses metadados e permitir novas opções de buscas aos mesmos.

Entretanto, vale ressaltar que a automatização citada nesse trabalho sobre a criação dos metadados das imagens é parcial, referindo-se somente às características de conteúdo visual da imagem. Parte desses metadados ainda são criados manualmente na forma de textos. Entretanto, muitas das informações das imagens que são descritas textualmente são objetivas (por exemplo, nome do autor, formato, data de criação, etc), e não causam tanta ambigüidade entre indexadores e usuários (a descrição de tais características não é tão problemática). Por outro lado, essa abordagem de unir textos e conteúdos visuais permite que consultas possam ser especificadas tanto através de palavras quanto através de uma imagem exemplo, oferecendo maiores vantagens e facilidades aos usuários e permitindo que sejam exploradas vantagens de ambos os métodos de consulta.

Ainda vale ressaltar que a abordagem de geração automática dos metadados das imagens e buscas por conteúdo a esses metadados, proposta nesse trabalho, é válida para outros tipos de recursos presentes nas Bibliotecas Digitais, tais como vídeo e som. Esses recursos também podem ter seus conteúdos extraídos automaticamente por algoritmos e inseridos nos metadados. Logo, as idéias discutidas nesse trabalho com relação à imagens podem ser adaptadas para esses outros tipos de recursos.

8.1 Dificuldades do Método Proposto

Como foi visto, metadados além de fornecer meios para a recuperação de recursos, também ajudam na compreensão do recurso encontrado. Um pequeno incômodo no uso de características visuais extraídas por algoritmos nos metadados das imagens é justamente em relação ao entendimento do conteúdo do recurso descrito. Tais características são geralmente valores numéricos (decimais ou binários) entendíveis, quase sempre, somente pelas máquinas, e não pelos humanos, diferentemente das características textuais.

Devido a essa desvantagem é interessante que as Bibliotecas Digitais mantenham metadados com elementos textuais e visuais. Assim, sistemas que promovem buscas aos recursos das bibliotecas, tais como os Provedores de Serviços, poderão combinar a recuperação textual e a recuperação por conteúdo. Dessa maneira, logo que os metadados das imagens recuperadas são apresentados ao usuário, esse tem a alternativa de ler os campos textuais, fazer um julgamento sobre o recurso, avaliar sua adequação para então decidir se ele é ou não adequado para seu propósito. Somente em caso satisfatório, o usuário precisará acessar o repositório no qual a imagem está armazenada (lembrando que somente os metadados estão disponíveis nos Provedores de Serviços, e não, os objetos digitais).

Por outro lado, uma dificuldade quando se trata de Sistemas de Recuperação de Imagens por Conteúdo é a necessidade de estabelecer um domínio específico para a implementação do sistema, justamente pela dificuldade de implementar módulos que trabalhem eficientemente para imagens de domínios variados, ou seja, de tipos diferentes.

Para que não seja necessário estabelecer um domínio específico aos Provedores de Serviços (a grande eficácia desses sistemas é justamente poder colher metadados de diferentes repositórios, independentemente do domínio desses repositórios), uma alternativa é que Provedores de Serviços implementem várias estratégias de recuperação baseada em conteúdo, para diferentes tipos das imagens. Assim, de acordo com o tipo da imagem fornecida pelo usuário, o Provedor de Serviços usaria o módulo de representação e recuperação adequado. No caso dos Provedores de Dados, o tipo da imagem poderia ser armazenado em seu metadado (em algum elemento Dublin Core ou em um novo elemento criado para tal tarefa) e deveria ser utilizado pelos Provedores de Serviços para a recuperação de imagens similares desse mesmo tipo.

Dificuldades adicionais para esse tipo de aplicação se fazem com relação ao problema de conseguir sistemas ágeis, rápidos e eficientes para efetuarem as buscas por conteúdo às imagens.

8.2 Trabalhos Futuros

Vários trabalhos poderão dar continuidade às idéias aqui propostas, tais como:

- Estudo sobre um modelo de extração de características e recuperação de imagens por conteúdo que poderia ser utilizado por Bibliotecas Digitais e Provedores de Serviços. Em outras palavras, um estudo sobre as melhores características visuais para representar os conteúdos das imagens, que deveriam ser inseridos nos metadados das imagens.

- Estudo sobre os principais algoritmos para o processamento das imagens, de acordo com o modelo definido.

- Estudo sobre como seria a implementação de um Sistema de Recuperação de Imagens por Conteúdo nos Provedores de Serviços (que atualmente só implementam a consulta por texto). Da mesma forma, como seria implementada a interface de busca desses provedores, e como o resultado das buscas seria mostrado ao usuário.

Referências Bibliográficas

[1] A. A. Araújo and S. J. F. Guimarães. Recuperação de informação visual com base no conteúdo em imagens e vídeos digitais. *RITA - Revista de Informática Teórica e Aplicada*, 7(2), 2000.

[2] A. G. R. Balan, A. J. M. Traina, C. Traina Jr., and P. M. A. Marques. Integrando textura e forma para a recuperação de imagem por conteúdo. *IX Congresso Brasileiro de Informática em Saúde - CBIS*, pages 6–6, 2004.

[3] J. M. Bueno. *Suporte à Recuperação de Imagens Médicas Baseada em Conteúdo através de Histogramas Métricos*. PhD thesis, Universidade de São Paulo, São Carlos – BR, 2001.

[4] E. Castro. *XML para a World Wide Web*. Editora Campos, 2001.

[5] CVIPtools. Em `http://www.ee.siue.edu/`, Último acesso: 22/06/2005.

[6] Instituto Brasileiro de Informação em Ciência e Tecnologia IBICT. Em `http://www.ibict.br/`, Último acesso: 20/03/2005.

[7] O. M. Filho and H. V. Neto. *Processamento Digital de Imagens*. Editora Brasport - Série Acadêmica, 1999.

[8] R. S. Filho, E. P. M. Souza, A. J. M. Traina, and C. Traina. Desmistificando o conceito de consultas por similaridade: A busca de novas aplicações na medicina. *II Workshop de Informática Médica*, 2002.

[9] Technical Advisory Service for Images. Metadata and digital images. *TASI*, June 2002.

[10] Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). Em `http://www.openarchives.org/OAI/openarchivesprotocol.html`, Último acesso: 03/04/2005.

[11] Representation for Understanding Images. Em `http://www.inf.ufpr.br/alexd`, Último acesso: 16/04/2005.

[12] S. S. Garcia. Metadados para documentação e recuperação de imagens. Master's thesis, Instituto Militar de Engenharia, Rio de Janeiro – BR, 1999.

[13] A. J. Gilliland-Swetland. Setting the stage: Introduction to metadata: Pathways to digital information. *In Baca M. (Ed.), Getty Information Institute*, 2002.

[14] R. C. Gonnzalez and R. E. Woods. *Digital Image Processing*. Addison-Wesley, 1993.

[15] A. Gupta and R. Jain. Visual information retrieval. *Communications of the ACM*, 40:71–79, 1999.

[16] Hypertext Transfer Protocol (HTTP). Em `http://www.w3.org/Protocols`, Último acesso: 01/02/2005.

[17] P. W. Huang and S. K. Dai. Image retrieval by texture similarity. *Pattern Recognition*, 36(3):665–679, March 2003.

[18] T. S. Huang, Y. Rui, and S. F. Chang. Image retrieval: Past, present and future. *Journal of Visual Communication and Image Representation*, 10:1–23, 1999.

[19] Namespaces in XML. Em `http://www.w3.org/TR/REC-xml-names`, Último acesso: 03/04/2005.

[20] Dublin Core Metadata Initiative. Em `http://www.dublincore.org/`, Último acesso: 01/02/2005.

[21] J. M. Jose, J. Furner, and D. J. Harper. Spacial querying for image retrieval: a user-oriented evaluation. In *Proceedings of 21st ACM SIGIR Conference on Research and development in information retrieval*, pages 232–240, Melbourne, Austrália, 1998.

[22] W. Meyer. Metodologias para classificação de texturas e consulta a base de imagens. Master's thesis, Instituto Militar de Engenharia, Rio de Janeiro – BR, 1997.

[23] H. Muller, N. Michoux, D. Bandon, and A. Geissbuhler. A review of content-based image retrieval systems in medical applications - clinical benefits and future directions. *Journals of Medical Informatics - Elsevier*, 2003.

[24] Open Archives Initiative (OAI). Em `http://www.openarchives.org`, Último acesso: 03/04/2005.

[25] OAIster. Em `http://www.oaister.org`, Último acesso: 10/04/2005.

[26] Marc Standards: Library of Congress Network Development, MARC, and Standard Office. Em `http://www.loc.gov/marc/`, Último acesso: 20/03/2005.

[27] M. Ortega, Y. Rui, K. Chakrabarti, K. Porkaew, S. Mehrotra, and T. S. Huang. Supporting ranked boolean similarity queries in mars. *IEEE Transactions on Knowledge and Data Engineering*, 10(6):905–924, 1998.

[28] H. Pedrini. *Apostila de Processamento de Imagens*. 2002.

[29] A. Pentland, R. W. Picard, and S. Sclaroff. Photobook: Content-based manipulation of image databases. *International Journal of Computer Vision*, 18(3):233–254, 1996.

[30] S. V. Pereira. Classificação e avaliação de sistemas e recuperação de imagens por conteúdo. Master's thesis, Universidade Federal do Paraná, Curitiba – BR, 2001.

[31] E. G. M. Petrakis, C. Faloutsos, and K. I. D. Lin. Imagemap: An image indexing method based on spatial similarity. *IEEE Transactions on Knowledge and Data Engineering*, 14(5):979–987, 2002.

[32] The Saxon XSLT Processor. Em `http://saxon.sourceforge.net/saxon6.5.2/index.html`, Último acesso: 03/04/2005.

[33] N. A. Rosa. Uma abordagem prática e eficiente de consultas por similaridade para suporte a diagnóstico por imagens. Master's thesis, Universidade de São Paulo, São Carlos – BR, 2002.

[34] Y. Ruy, T. S. Huang, and S. Mehtotra. Content-based image retrieval with relevance feedback in mars. In *Proceedings of IEEE International conference on Image Processing*, pages 815–818, Santa Barbara, California, EUA, 1997.

[35] M. Safar, C. Shahabi, and X. Sun. Image retrieval by shape: A comparative study. New York, 1999.

[36] XML Schema. Em `http://www.w3.org/XML/Schema`, Último acesso: 03/04/2005.

[37] Dublin Core Metadata Element Set. Em `http://dublincore.org/documents/dces/`, Último acesso: 03/04/2005.

[38] M. Severich. Uma ferramenta de processamento de imagens para o sistema footscan. Master's thesis, Universidade Federal do Paraná, Curitiba – BR, 2002.

[39] G. Sheikholeslami, W. Chang, and A. Zhang. Semquery: Semantic clustering and querying on heterogeneous features for visual data. *IEEE Transactions on Knowledge and Data Engineering*, 14(5):988–1002, September 2002.

[40] OAI Tech. Em `http://www.openarchives.org/organization/index.html`, Último acesso: 03/04/2005.

[41] A. J. M. Traina. *Suporte à Visualização de Consultas por Similaridade em Imagens Médicas através de Estrutura de Indexação Métrica*. PhD thesis, Universidade de São Paulo, São Carlos – BR, 2001.

[42] DCMI Type Vocabulary. Em `http://dublincore.org/documents/dcmi-type-vocabulary/`, Último acesso: 03/04/2005.

[43] D. J. Waters. What are digital libraries? *Digital Library Information Resources in Berkeley Digital Library SunSite, CLIR Issues*, (4), 1995.

[44] Extensible Markup Language (XML). Em `http://www.w3.org/XML`, Último acesso: 12/12/2004.

[45] XSL Transformations (XSLT). Em `http://www.w3.org/TR/xslt`, Último acesso: 03/04/2005.

[46] D. Zhang and G. Lu. Content-based shape retrieval using different shape descriptors: A comparative study. pages 317–320, Tokio, Japan, August 2001.

Anexo I

Esquemas XML e o Suporte para Múltiplos Formatos de Metadados no Protocolo OAI-PMH

O protocolo OAI-PMH utiliza Esquemas XML (*XML Schemas*) para definir o formato dos metadados dos registros a serem colhidos. O esquema XML desenvolvido pela OAI para validar os metadados a serem colhidos é chamado oai-dc. O oai-dc é um formato de metadados simples, cujos elementos são baseados no Dublin Core *Unqualified*. Este esquema é usado como o formato de metadados mínimo para interoperabilidade exigido pelo protocolo OAI-PMH. Isto quer dizer que, a menos que determinado Provedor de Dados defina seu próprio formato de metadados através dos Esquemas XML e valide-o junto à OAI, os metadados dos seus registros deverão estar de acordo com o formato definido no oai-dc.

O oai-dc define um esquema recipiente (*container schema*) específico com a OAI e que fica hospedado em seu site, e importa um esquema genérico DCMES (Conjunto de Elementos dos Metadados Dublin Core), hospedado no site da DCMI (Iniciativa de Metadados Dublin Core). O mesmo modelo poderia ser usado para um esquema Dublin Core *Qualified*, isto é, um esquema recipiente hospedado pela OAI que referencia o esquema genérico Dublin Core *Qualified*, hospedado pela DCMI.

A documentação do OAI-PMH descreve o uso dos Esquemas XML para outros formatos de metadados, fornecendo esquemas adicionais para:

- rcf1807 (para o formato de metadados RFC 1807)
- marc21 (recomendado para metadados MARC21, fornecido pela *Library of Congress*)
- oai-marc (para o formato de metadados MARC)

oai-dc - Exemplo de um Registro

Abaixo se segue o exemplo de um registro que segue o esquema oai-dc, visto através da ferramenta Repository Explorer [24]. Este é um trecho de um registro retornado como resposta à requisição do verbo

GetRecord (um dos seis verbos definidos pelo protocolo OAI-PMH para validar e fazer requisições aos repositórios):

```
<?xml version="1.0" encoding="UTF-8"?> <OAI-PH
xmlns="http://www.openarchives.org/OAI/2.0/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
      xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
      http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
 <responseDate>2003-03-15T16:16:51+01:00</responseDate>
 <request verb="GetRecord" metadataPrefix="oai_dc"
 identifier="oai:HUBerlin.de:3000476">http://edoc.hu-berlin.de/OAI-2.0</request>
 <GetRecord>
 <record>
 <header>
 <identifier>oai:HUBerlin.de:3000476</identifier>
 <datestamp>1997-07-18</datestamp>
 <setSpec>pub-type</setSpec>
 </header>
 <metadata>
 <oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
            xmlns:dc="http://purl.org/dc/elements/1.1/"
            xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
            xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
            http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
 <dc:title>Melanchthon in seiner Zeit. In: Philipp Melanchthon
 1497-1997</dc:title>
 <dc:creator>Selge, Kurt-Victor</dc:creator>
 ...
```

Há três informações importantes no trecho do registro mostrado acima:

- O namespace para o formato: oai-dc:xmlns:oai-dc="http://www.openarchives.org/OAI/2.0/oai-dc/"
- O namespace para os elementos DCMES:xmlns:dc="http://purl.org/dc/elements/1.1/"
- Esquema recipiente associado com o namespace oai-dc: xsi:schemaLocation= "http://www.openarchives.org/OAI/2.0/oai-dc/ http://www.openarchives.org/OAI/2.0/oai-dc.xsd"

Adicionando Novos Elementos quando o oai-dc não é suficiente

Criar um novo esquema estendendo o oai-dc, através da adição de novos elementos, envolve as seguintes tarefas:

1. Criar um nome para o novo esquema.
2. Criar namespaces.
3. Criar o esquema para os novos elementos.
4. Criar um esquema recipiente (*container schema*).
5. Validar o esquema/registros.
6. Informar aos verbos do OAI-PMH sobre o novo formato.
7. Testar se o esquema funciona e se é válido.

Passo 1: Criar um nome para o novo esquema

O novo formato de metadados precisa de um nome. Para o formato proposto por esse trabalho, foi escolhido o nome "ime-dc", seguindo o nome do formato da OAI, "oai-dc", como convenção ("ime" é uma abreviação do termo "imagens médicas"). Entretanto, o nome poderia ser qualquer um, não existindo nenhuma regra para tal.

Passo 2: Criar namespaces

São necessários dois namespaces:

- Um namespace para o novo formato (ime-dc), que combina tantos os elementos Dublin Core quanto os outros elementos quaisquer;
- Um namespace para os novos elementos de metadados (tal como o elemento "perimeter", por exemplo) criados para o novo formato.

Namespaces são declarados como URIs. São usados:

- http://www.inf.ufpr.br/ beatriz/metadados/imagensmédicas/ime-dc/ (para o primeiro namespace)
- http://www.inf.ufpr.br/ beatriz/metadados/imagensmédicas/ime-elementos/ (para o segundo namespace)

O uso de PURL para os namespaces segue o uso da DCMI, mas isso não é obrigatório. Entretanto, ambas as URIs dos namespaces devem ser controladas por um responsável para garantir que sejam únicas e para prevenir reuso no futuro. As URIs dos namespaces não devem, necessariamente, apontar para algum lugar.

Passo 3: Criar o esquema para os novos elementos

Um esquema XML deve ser criado para os novos elementos. Por exemplo:

http://www.inf.ufpr.br/ beatriz/metadados/imagensmédicas/ime-dc/15112004/ime-elementos.xsd

Geralmente são criados diretórios nomeados com as datas de criação do esquema, para tornar mais fácil referenciar o esquema, tanto o atual, quanto os mais antigos. O esquema para os novos elementos define tais elementos e adiciona-os ao grupo dc:any. Ele também define um novo tipo recipiente "ime-elementos:elementContainer".

Passo 4: Criar um esquema recipiente (*container schema*)

É necessário também criar um esquema recipiente para o formato ime-dc. Por exemplo:

http://www.inf.ufpr.br/ beatriz/metadados/imagensmédicas/ime-dc/15112004/ime-dc.xsd

Aqui novamente é usado o diretório nomeado com a data da criação desse esquema. Esse esquema simplesmente importa o esquema ime-elementos e então define um elemento recipiente "ime-dc" do tipo ime-elementos:elementContainer.

Passo 5: Validar o esquema/registros

Para validar os registros através do novo esquema, são criados alguns registros para os testes, incluindo todos os elementos criados. Nesse momento, tais registros podem ser validados junto ao validador de Esquemas XML, disponível no endereço http://www.w3.org/2001/03/webdata/xsv/.

Passo 6: Informar aos verbos do OAI-PMH sobre o novo formato

Cinco dos seis verbos do protocolo OAI-PMH usados para colher metadados precisam ser informados sobre o novo formato. São eles: ListMetadataFormats, ListSets, ListIdentifiers, ListRecords e GetRecord. Para tanto, é necessário modificar o software do repositório (código fonte e/ou arquivos de configuração) onde o novo formato será usado, de forma que o nome do novo formato "ime-dc" seja aceito com um "MetadataPrefix". Isso é feito modificando as repostas do repositório a esses verbos. Assim, os repositórios retornarão os registros formatados de acordo com o novo esquema, quando requisitados por Provedores de Serviços.

Abaixo se segue um exemplo com o verbo ListMetadataFormats:

```
...
<metadataFormat>
<metadataPrefix>ime_dc</metadataPrefix>
<schema>
http://www.inf.ufpr.br/~beatriz/metadados/imagensmédicas/ime_dc/15112004/ime_elementos.xsd
</schema>
<metadataNamespace>
http://www.inf.ufpr.br/~beatriz/metadados/imagensmédicas/ime_dc/
```

```
</metadataNamespace>
</metadataFormat> ...
```

Passo 7: Testar se o esquema funciona e se é válido

Nesse ultimo passo é utilizada a ferramenta Repository Explorer [24] para testar o novo formato. Para tanto, deve-se indicar a URL da interface OAI do repositório. Por exemplo:

http://www.inf.ufpr.br/ beatriz/metadados/oai/nph-oai2.cgi

Esse teste deve garantir que:

- Todas as requisições funcionam com o novo "metadataPrefix";
- O formato oai-dc ainda funciona;
- Os registros são retornados corretamente para cada formato;
- As respostas são validadas corretamente.

Quando todas essas condições são satisfeitas, tem-se um novo formato.