Universidade Federal de Pernambuco
Centro de Informática

Pós-graduação em Ciência da Computação

Clio-i: Interoperabilidade entre repositórios digitais utilizando o protocolo OAI-PMH

Marcos José de Menezes Cardoso Junior

Dissertação de Mestrado

Recife
27 de fevereiro de 2007

Universidade Federal de Pernambuco
Centro de Informática

Marcos José de Menezes Cardoso Junior

Clio-i: Interoperabilidade entre repositórios digitais utilizando o protocolo OAI-PMH

Trabalho apresentado ao Programa de Pós-graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

Orientadora: *Profa. Dra. Flávia de Almeida Barros*
Co-orientador: *Prof. Dr. Ricardo Prudêncio*

Recife
27 de fevereiro de 2007

A Painho e Mainha.

Agradecimentos

Agradeço a todos que, direta ou indiretamente, contribuiram para o desenvolvimento deste trabalho. Em especial a alguns:

- Aos professores Flávia Barros e Ricardo Prudêncio, pela forma paciente e dedicada que me orientaram.
- Ao professor e coordenador do laboratório Liber, Marcos Galindo Lima, que me orienta desde a época da graduação e grande resposável pela idealização do trabalho.
- À toda minha família, em especial aos meus pais, Marcos e Carminha, e minha irmãs, Fabiana e Tatiana, que sempre me deram apoio e suporte na época do mestrado.
- À minha afilhada Clara, que me concedeu a alegria de ser padrinho.
- A todos os meus amigos do Centro de Infórmatica: Pereira, Capiau, Dedeco, Gandhi, Glauber, Sorriso, Presidente, Pimbolim, Leozinho, Aline, Anchieta, Lelo, Victor, Dominoni, Heitor, Rodrigão, Cabelo de Veludo, Taciana, Victor, Juliana, Cabelinho, Narinha, Brasil, André, Mazza, Babita, Jarbinhas e tantos outros que sempre me ajudaram na época da graduação e do mestrado.
- Ao Laboratório Liber, que me ajudou de todas as formas na realização do trabalho.
- À AVCin (que um dia vai fazer sucesso), ao D'Breck (que já é um sucesso) e aos Infames (que nunca vai fazer sucesso).
- Ao Sport Club do Recife, time que me deu tantas alegrias no período do mestrado, e continuará dando! PST!!!
- E, principalmente, a Deus, por ter me colocado todas essas pessoas em meu caminho.

A seta partiu
Do arco retesado
E com grande violência
Foi atingir o mísero cervo
Em disparada
A mão segura e firme do caçador
A precisão da pontaria
A destreza do arremesso
Foram um princípio eficiente
Para um final esperado.
Assim sejam as nossas atitudes
Firmes, resolutas, precisas
Formadas com toda a retidão
Para que o nosso alvo
Seja sempre atingido com honra.

—EDUARDO CABRAL DE MELO

Resumo

O interesse na criação de Bibliotecas Digitais cresceu significativamente a partir do surgimento e da disseminação da Web, que trouxe consigo a necessidade de ferramentas que facilitassem a publicação, gerenciamento e a recuperação da informação digital. Atualmente, podemos observar uma ampla gama de Bibliotecas Digitais, que se caracterizam como serviços que gerenciam e disponibilizam documentos digitais, de forma mais estruturada do que convencionalmente se observa na Web. Nesses serviços, em geral, os documentos são descritos com metadados, recuperados através de ferramentas de busca estruturada e visualizados em interfaces apropriadas. Uma das limitações de grande parte das Bibliotecas Digitais existentes é a ausência de mecanismos de integração de dados, de maneira a fornecer ao usuário, acesso unificado e transparente aos repositórios gerenciados por diferentes serviços. Esse problema é conhecido na literatura como o problema da Interoperabilidade de Bibliotecas Digitais. Dentro desse contexto, desenvolvemos o Clio-i, uma arquitetura de sistema para gerenciamento de Bibliotecas Digitais. Essa arquitetura apresenta características desejáveis como um módulo de recuperação de documentos, um visualizador de documentos e um módulo de interoperabilidade entre repositórios digitais. Para prover o mecanismo de integração, foi implementada no Clio-i uma extensão do protocolo OAI-PMH (Open Archives Initiative-Protocol for Metadata Harvesting), que é um padrão internacional para interoperabilidade de repositórios digitais. O módulo integrador do Clio-i é composto de duas partes: (1) o Clio-i Data Provider, responsável por exportar os metadados dos documentos gerenciados localmente, de acordo com os padrões estabelecidos no OAI-PMH; e (2) o Clio-i Service Provider, que realiza a coleta de informações a partir de qualquer provedor de dados remoto baseado em OAI-PMH. O protótipo de sistema implementado foi validado em dois estudos de caso, operacionalizando centenas de milhares de registros e efetivando a extensão do protocolo OAI-PMH adotada.

Palavras-chave: Bibliotecas Digitais, Integração de Dados, OAI-PMH

Abstract

O interesse na criação de Bibliotecas Digitais cresceu significativamente a partir do surgimento e da disseminação da Web, que trouxe consigo a necessidade de ferramentas que facilitassem a publicação, gerenciamento e a recuperação da informação digital. Atualmente, podemos observar uma ampla gama de Bibliotecas Digitais, que se caracterizam como serviços que gerenciam e disponibilizam documentos digitais, de forma mais estruturada do que convencionalmente se observa na Web. Nesses serviços, em geral, os documentos são descritos com metadados, recuperados através de ferramentas de busca estruturada e visualizados em interfaces apropriadas. Uma das limitações de grande parte das Bibliotecas Digitais existentes é a ausência de mecanismos de integração de dados, de maneira a fornecer ao usuário, acesso unificado e transparente aos repositórios gerenciados por diferentes serviços. Esse problema é conhecido na literatura como o problema da Interoperabilidade de Bibliotecas Digitais. Dentro desse contexto, desenvolvemos o Clio-i, uma arquitetura de sistema para gerenciamento de Bibliotecas Digitais. Essa arquitetura apresenta características desejáveis como um módulo de recuperação de documentos, um visualizador de documentos e um módulo de interoperabilidade entre repositórios digitais. Para prover o mecanismo de integração, foi implementada no Clio-i uma extensão do protocolo OAI-PMH (Open Archives Initiative-Protocol for Metadata Harvesting), que é um padrão internacional para interoperabilidade de repositórios digitais. O módulo integrador do Clio-i é composto de duas partes: (1) o Clio-i Data Provider, responsável por exportar os metadados dos documentos gerenciados localmente, de acordo com os padrões estabelecidos no OAI-PMH; e (2) o Clio-i Service Provider, que realiza a coleta de informações a partir de qualquer provedor de dados remoto baseado em OAI-PMH. O protótipo de sistema implementado foi validado em dois estudos de caso, operacionalizando centenas de milhares de registros e efetivando a extensão do protocolo OAI-PMH adotada.

Keywords: Digital Libraries, Data Integration, OAI-PMH

Sumário

Lista de Figuras

Lista de Tabelas

CAPÍTULO 1

Introdução

A informação é a única matéria-prima que se reproduz quando é disseminada.

—VAN BENTHEM

O rápido avanço da Internet, em especial da Web, levou à possibilidade de acesso a uma gama de serviços informacionais, dentre os quais destacamos as Bibliotecas Digitais. Esse serviço é um dos tipos mais avançados e complexos de sistemas de informação, por envolver, dentre outras coisas, busca estrutura e navegação de documentos, preservação do documento, serviços de informação multimídia e disseminação seletiva da informação [FM98].

A partir da segunda metade da década de 1990, as pesquisas no campo das Bibliotecas Digitais tiveram grande crescimento [WMBB00]. Essa pesquisa proporcionou o desenvolvimento de Bibliotecas Digitais atendendo a comunidades com interesses de informação específicos, além de ferramentas genéricas que dão suporte a criação de novos serviços. Contudo, poucas ferramentas dispõem de algum mecanismo de integração, de forma a oferecer acesso simplificado a documentos disponibilizados por Bibliotecas Digitais distintas.

Dentro do contexto acima, investigamos o problema da Interoperabilidade de Bibliotecas Digitas, e propomos o Clio-i, uma arquitetura que oferece aos usuários mecanismos desejáveis a uma Biblioteca Digital, destacando um módulo de interoperabilidade de repositórios digitais.

1.1 Interoperabilidade de Bibliotecas Digitais e o OAI-PMH

No fim da década de 90 já existiam algumas Bibliotecas Digitais na Internet, distribuídas em diversas áreas. Algumas delas, inclusive, realizam a interoperabilidade de seus dados com outras instituições. Contudo, cada repositório implementava um protocolo próprio, trazendo dificuldades no compartilhamento de seus metadados entre servidores distintos. Essa disseminação estava sendo prejudicada, devido ao fato de usuários encontrarem diferentes interfaces, tornando o processo de busca mais difícil. Além do mais, não havia uma forma automática de compartilhar os dados [LdS01].

Dessa maneira, muitos protocolos foram criados com a finalidade de sanar tais problemas, dentre os quais podemos destacar o Z39.50 [Lyn97], SDLIP [Ubi], NCSTRL [Lei98] e o OAI [OAIb]. O último, o *Open Archives Initiative*, difere-se dos demais devido a sua praticidade e eficiência na área e por isso foi o padrão escolhido para os nossos estudos.

O *Open Archives Initiative* (OAI) é uma organização internacional que promove padrões para permitir a interoperabilidade entre repositórios digitais na Internet [Brab] . Um resultado

importante dessa iniciativa foi a formulação do protocolo de integração de Bibliotecas Digitais, o *Open Archives Initiative Protocol for Metadata Harvesting* (OAI-PMH). Uma das grandes vantagens do protocolo é o seu baixo custo de implementação por se basear em tecnologias já bastantes difundidas e padronizadas internacionalmente, como o HTML, o XML, e padrão de metadados Dublin Core.

A base do OAI é o protocolo OAI-PMH, que faz com que os participantes da iniciativa possam compartilhar seus metadados através de regras bastante claras e simples. Além destas regras, há dois grupos "participantes": os Provedores de Dados e os Provedores de Serviços. Os Provedores de Dados são repositórios que armazenam os recursos digitais e implementam o protocolo OAI-PMH como forma de expor os metadados de seus documentos. Já os Provedores de Serviço utilizam o protocolo para coletar os metadados, armazená-los e utilizá-los em algum serviço oferecido [LdS01].

Apesar de ser um protocolo bastante difundido pela Internet, identificamos algumas necessidades que não são supridas pelo OAI-PMH (e.g.: exportação do recurso eletrônico, coleta de metadados através de algumas palavras-chave, etc.). Desta maneira, realizamos uma extensão ao protocolo OAI-PMH e a aplicamos no Clio-i.

1.2 A Arquitetura do Clio-i

Uma vasta pesquisa foi realizada, analisando diferentes sistemas de Bibliotecas Digitais. A partir das funcionalidades e limitações identificadas nesses sistemas, foram definidos critérios de avaliação para sistemas de Bibliotecas Digitais. A partir desses critérios, através de uma metodologia de prototipação evolutiva para o desenvolvimento de sistemas, chegamos na proposta atual do nosso trabalho, a arquitetura do sistema Clio-i, para gerenciamento e acesso de repositórios de documentos digitais.

A arquitetura proposta se diferencia de outros sistemas do gênero por oferecer funcionalidades como gerenciamento de acervo multimídia, mecanismo automático de indexação e busca de documentos, uma interface para a visualização dos documentos, e a integração de dados. Para isso, o sistema foi desenvolvido a partir de uma arquitetura que contempla quatro componentes principais:

- **Recuperação de Informação:** Tem a finalidade de realizar a pesquisa em todos os documentos inseridos na base através dos metadados (e.g. titulo, autor, resumo, etc.) que o descrevem. Esta resposta à consulta é retornada ao usuário em ordem de relevância, provendo ao usuário fácil acesso às informações desejáveis.

- **Visualização de Documentos:** O usuário pode visualizar um documento específico de sua escolha, através deste módulo, trabalhando com arquivos do tipo áudio, imagem, vídeo e texto.

- **Clio-i Data Provider:** Esse componente exporta o Banco de Metadados do sistema em um formato específico de acordo com o protocolo OAI-PMH. Isso possibilita que os metadados dos documentos sejam colhidos por outros provedores de serviços externos que adotam o protocolo OAI-PMH e integrados a Bibliotecas Digitais externas.

- **Clio-i Service Provider:** Realiza a coleta de metadados referentes a documentos disponibilizados por Provedores de Dados OAI, inclusive de Clio-i Data Providers externos.

A arquitetura do Clio-i foi avaliada em um protótipo implementado em PHP[PHPa], uma linguagem adequada para criação de páginas dinâmicas na Web. O banco de dados escolhido para o armazenamento dos metadados foi o MySQL[MySb]. O módulo de Recuperação de Informação foi implementado utilizando o *MySQL Full-Text Search*[MySa], que permite, por exemplo, casamento de palavras-chave e expressões de busca booleana e ordenação dos resultados usando um critério de relevância. O módulo Visualizador de Documentos foi implementado utilizando a tecnologia DHTML (Dynamic HTML)[GS03], o que permitiu a implementação de operações realizadas nos documentos (e.g. zoom in, zoom out, girar, inverter,...) de forma online e com um bom tempo de resposta.

Para o perfeito funcionamento do Clio-i Data Provider, alguns componentes tiveram que ser implementados, dentre eles, um parser, com a função de validar as requisições OAI e o controle de fluxo, responsável por exportar as informações em porções pré-definidas.

Para realizar a coleta de diferentes repositórios OAI, o Clio-i Service Provider foi implementado utilizando uma biblioteca nativa do PHP chamada SimpleXML[PHPb]. Esta biblioteca visa integrar, de maneira simples, XML no PHP, trabalhando numa estrutura de objetos.

Todo o projeto do Clio-i foi implementado e os resultados foram aplicados em dois estudos de caso. O primeiro estudo de caso apresentado corresponde ao Integrador de Repositórios Científicos. A proposta desse projeto era reunir, em uma única base, diversos metadados correspondentes a arquivos eletrônicos científicos de todo o mundo. Para concretizarmos o projeto, foi escolhido dezenove Provedores de Dados oficiais da OAI. A coleta desses repositórios reuniu mais de centro e trinta mil registros, que são pesquisáveis através do módulo de Recuperação de Informação.

O segundo estudo caso trata-se da reunião de documentos presentes em três Clio-i Data Providers, cada um com diferentes projetos:

- **Acervo Digital FUNDAJ:** Documentos no formato imagem do acervo da Fundação Joaquim Nabuco.

- **Holandeses na Bahia:** Base composta por documentos do tipo texto, sobre a conquista holandesa de Salvador no Brasil.

- **Escrito nas Estrelas:** Trata-se de uma exposição sobre o Brasil na época do desenvolvimento da aviação no início do século XX, cujos arquivos estão no formato áudio e vídeo.

Neste segundo estudo de caso demonstramos a extensão do protocolo OAI-PMH proposta, coletando não só os metadados de uma base, mas como também os seus recursos eletrônicos em formato de imagem, vídeo, texto e áudio.

1.3 Organização da Dissertação

Além da Introdução, esta dissertação conta com mais seis capítulos, como se segue:

- **Capítulo 2 - Bibliotecas Digitais e Integração:** Esse capítulo apresenta diversas definições sobre Bibliotecas Digitais, além de explicar o que são metadados e suas aplicações, especificando o padrão de metadados Dublin Core. O capítulo também realiza comentários sobre integração de dados, informando abordagens e arquiteturas. Por fim, define critérios de avaliação para Bibliotecas Digitais.

- **Capítulo 3 - Open Archives Initiative - OAI:** O capítulo explica tece comentários sobre a iniciativa do *Open Archives*, explicando detalhadamente o seu protocolo, o OAI-PMH. Também define critérios de avaliação para Bibliotecas Digitais que adotam o protocolo.

- **Capítulo 4 - O Clio-i:** Esse capítulo apresenta a definição do sistema proposto, especificando os seus requisitos, descrevendo com detalhes sobre a arquitetura do software e os seus três componentes principais. Ainda explica sobre a extensão do protocolo OAI-PMH adotada no Clio-i.

- **Capítulo 5 - Protótipo e Testes:** O capítulo 5 mostra todos os detalhes de implementação do Clio-i, apresentando a metodologia de construção utilizada, a estrutura da base de dados e os módulos processadores. Ainda realiza dois tipos de testes: funcional e de usabilidade.

- **Capítulo 6 - Estudos de Caso:** Este capítulo tem a finalidade de apresentar dois estudos de caso utilizados para aplicar todas as funcionalidades do Clio-i. No primeiro deles, é testada a sua eficiência, operacionalizando centenas de milhares de registros vindos de diversas bases na Internet. O objetivo principal do segundo estudo de caso é mostrar, principalmente, a extensão do protocolo OAI-PMH em um projeto real.

- **Capítulo 7 - Conclusões:** Realiza as conclusões do trabalho apresentado, ressaltando as suas contribuições. Apresentamos ainda trabalhos futuros relacionados que darão continuidade ao sistema Clio-i.

Além desses capítulos, este documento traz ainda quatro apêndices:

- **Apêndice A - Relatório de Avaliação dos Testes Funcionais:** Tem o objetivo de apresentar os testes funcionais que foram realizados no Clio-i, apresentando os resultados e modificações realizadas.

- **Apêndice B - Plano de Teste de Usabilidade:** Descreve com detalhes sobre o Teste de Usabilidade realizado no Clio-i, a metodologia utilizada e as medidas de avaliação sobre o sistema.

- **Apêndice C - Lista de Tarefas:** Apresenta com detalhes as vinte e duas tarefas que foram realizadas no Teste de Usabilidade pelos participantes.

- **Apêndice D - Questionário de Avaliação do Sistema pelo Participante:** Este apêndice tem o objetivo de colher informações sobre a opinião dos participantes do Teste de Usabilidade realizado no Clio-i.

CAPÍTULO 2

Bibliotecas Digitais e Integração

Se transmito a você uma informação, não a perco, e se a utilizo, não a destruo.
—PIERRE LÉVY

É bastante comum a noção de Bibliotecas Digitais (ou Bibliotecas Virtuais) como extensões de bibliotecas tradicionais encontradas em escolas, universidades ou museus. Tal senso é formado apenas sobre aquisição e digitalização de dados. A captura e passagem para o formato digital da informação são chaves para o entendimento, contudo Bibliotecas Digitais são mais do que simples coleções digitais [BYRN99]. Assim sendo, é de fundamental importância entendermos alguns conceitos a respeito deste importante assunto.

Neste capítulo, apresentaremos uma revisão sobre Bibliotecas Digitais, abordaremos o conceito e alguns padrões de Metadados, citaremos assuntos relevantes sobre integração de dados no tema proposto para o trabalho e, por fim, apresetaremos alguns critérios adotados para avaliação de Bibliotecas Digitais.

2.1 Bibliotecas Digitais

Com o avanço da Internet, a sociedade passou a ter acesso a uma gama de recursos informacionais através de e-mails, listas de discussão, artigos em revistas eletrônicas, informações comerciais, culturais, artísticas e Bibliotecas Digitais [RH02]. Especificamente sobre este último aspecto, [FM98] afirma que uma Biblioteca Digital envolve "suporte de forma colaborativa, preservação de documento digital, gerenciamento de base de dados distribuída, hipertexto, filtros e recuperação de informação, módulos de instrução, gerenciamento de direitos autorais, serviços de informação multimídia, serviços de referência e respostas às questões enviadas, busca de recursos, e disseminação seletiva", sendo, desta maneira, uma das formas mais avançadas e complexas de sistemas de informação.

Desde que o assunto passou a ser comumente referenciado em conferências e workshops, há uma contínua discussão de como definir uma Biblioteca Digital [FM98]. Muitos autores, inclusive, afirmam que a definição sobre Bibliotecas Digitais ainda encontra-se em fase de maturação [RH02]. Na verdade, todas elas apontam a serviços ou características particulares que uma Biblioteca Digital realiza em um determinado repositório para um público-alvo específico. Segundo [Atk98]: "o conceito de Biblioteca Digital está na analogia com um lugar onde se encontra um repositório contendo uma coleção organizada de publicações (que possam ser impressos) e outros artefatos físicos, combinados com sistemas e serviços que facilitem o

acesso físico, intelectual, e disponível por longo tempo".

Diferentes definições de Bibliotecas Digitais podem ser encontradas ainda na literatura. Em uma das primeiras definições, temos: "Uma biblioteca que mantêm toda ou uma parte substancial de sua coleção numa forma processável pelo computador como uma alternativa, suplemento ou complemento à forma impressa tradicional e material em microfilme, que, atualmente, domina os acervos bibliográficos"[Saf95].

Uma outra definição que consideramos mais atual e ampla do que a anterior é definida pela Digital Library Federation [DLF]: "Bibliotecas digitais são organizações que fornecem recursos para selecionar, estruturar, oferecer acesso intelectual, distribuir, preservar a integridade e garantir a permanência das coleções digitais, de tal forma que elas estejam disponíveis para uma ou várias comunidades".

Apesar de mais completo, a definição acima não atende ainda ao uso de diferentes tipos de arquivos digitais, comumente utilizados hoje em dia. A definição de [CDD+96] preenche essa lacuna: "Uma coleção organizada de dados multimídia com métodos de gerenciamento da informação, que representa os dados como informação útil e conhecimento para o povo numa variedade de contextos sociais e organizacionais".

Outras definições poderiam ser citadas, contudo outros aspectos sobre Bibliotecas Digitais devem ser destacados. Inicialmente, destacamos que a maioria das Bibliotecas Digitais disponíveis apresenta dificuldades quanto ao acesso das informações nelas contidas. Ora a pesquisa dá-se de forma complexa, ora os resultados obtidos não são relevantes. Há casos em que são retornados documentos de interesse do pesquisador, mas o sistema disponibiliza apenas sua referência, sendo esse resultado da pesquisa muitas vezes inútil e a ida a uma biblioteca tradicional indispensável [Car05].

Ainda, com o avanço das técnicas de Recuperação de Informação [BYRN99], poderíamos criar Bibliotecas Digitais com poderosos sistemas de busca e resultados os mais relevantes possíveis, apresentados em um tempo de resposta aceitável. Além disso, as instituições detentoras desses acervos virtuais começaram a vislumbrar novas possibilidades. A principal delas foi perceber as vantagens de compartilhar seus diferentes acervos temáticos com outras instituições. Isso facilitaria o acesso dos pesquisadores a documentos digitais em diferentes repositórios, através de uma única interface.

Finalmente, além das definições de diferentes autores a respeito de Biblioteca Digital, devemos abordar os aspectos citados anteriormente. Com isto, temos uma Biblioteca Digital como *sistemas que provêem acesso à informação de forma clara, objetiva, relevante, completa e integrada com outros acervos digitais*. Acreditamos que essa definição atinge os aspectos essenciais das Bibliotecas Digitais.

2.2 Metadados

Com o avanço da Internet, o volume de informações disponíveis cresceu substancialmente, causando dificuldades para a disseminação da informação. Dentre essas dificuldades podemos citar o grande número de detentores de informações e seus altos graus de autonomia e a falta de uma estrutura para acolher esses dados [LPV04]. Com isso, o desenvolvimento de padrões que descrevam essas informações de forma mais estruturada torna-se imprescindível para aquelas

instituições que desejam disponibilizar os seus dados, ou, focado no presente trabalho, para disseminar os seus acervos digitais [Car05].

Como parte da solução do problema descrito acima, novos formatos para estruturar e disponibilizar a informação eletrônica estão sendo desenvolver para o acesso através da Internet. Esses formatos estão sendo designados como metadados e já constituem um grande conjunto de normas aplicáveis à gestão da informação digital [RH02].

Muitos são os conceitos encontrados sobre o assunto. Segundo a definição de [Tro98]: "Metadado é a descrição do dado, do ambiente onde ele reside, como ele é manipulado e para onde ele é distribuído". Ou seja, trata-se de informações estruturadas sobre os recursos presentes em um repositório de dados. Tais recursos podem ser imagens, livros, músicas, artigos científicos, documentos históricos, dentre muitos outros.

Imaginemos, por exemplo, recursos como um livro, uma obra de arte, uma fotografia e uma página na Web. Sabemos que um livro e uma obra de arte possuem um autor; uma foto possui um fotógrafo que a reproduziu; e uma página na web certamente um *webmaster*. Esses dados (ou metadados) poderiam ser vistos como um elemento padronizado dentro de um mesmo repositório. Por exemplo, um elemento creator poderia fazer referência aos autores, fotógrafos e *webmasters* de cada recurso específico. Na figura 2.1, podemos visualizar os diferentes metadados criados para cada recurso.

Figura 2.1 Diferentes recursos associados aos seus metadados.

Além da padronização dos recursos de um repositório (uma Biblioteca Digital, por exem-

plo), os metadados também possuem a função de permitir a interoperabilidade entre diferentes fontes de dados. Desta forma, a padronização dos recursos através de um padrão de metadados pode melhorar substancialmente a coleta da informação por outros sistemas.

O rápido crescimento da Internet também trouxe consigo uma ploriferação de padrões de metadados, cada um construído para atender uma comunidade, tipos de materiais e necessidades de projetos específicos [ZC06]. Usuários que realizam consultas em uma Biblioteca Digital não devem ter a obrigação de entender os métodos usados para descrever e representar o conteúdo de uma coleção digital [Ten01]. Entretanto, pela complexidade e extensão de alguns padrões de metadados, a sua adoção pode indicar desafios para a descrição e o gerenciamento de recursos de um repositório, inclusive pelos próprios desenvolvedores da aplicação. Dentre os padrões de metadados, podemos destacar alguns como o Dublin Core [Corb], o MARC [oC], o IEEE Learning Object Metadata [IEE], e o IMS [MMGG01].

Dos padrões citados acima, o Dublin Core (DC) e o MARC são os mais utilizados hoje em dia, conforme mostra o estudo realizado em [KPT03]. Esse estudo contemplou 227 instituições que possuem algum tipo de acervo digital dentre bibliotecas acadêmicas e públicas, museus, arquivos históricos e outras. A figura 2.2 resume os resultados do estudo.

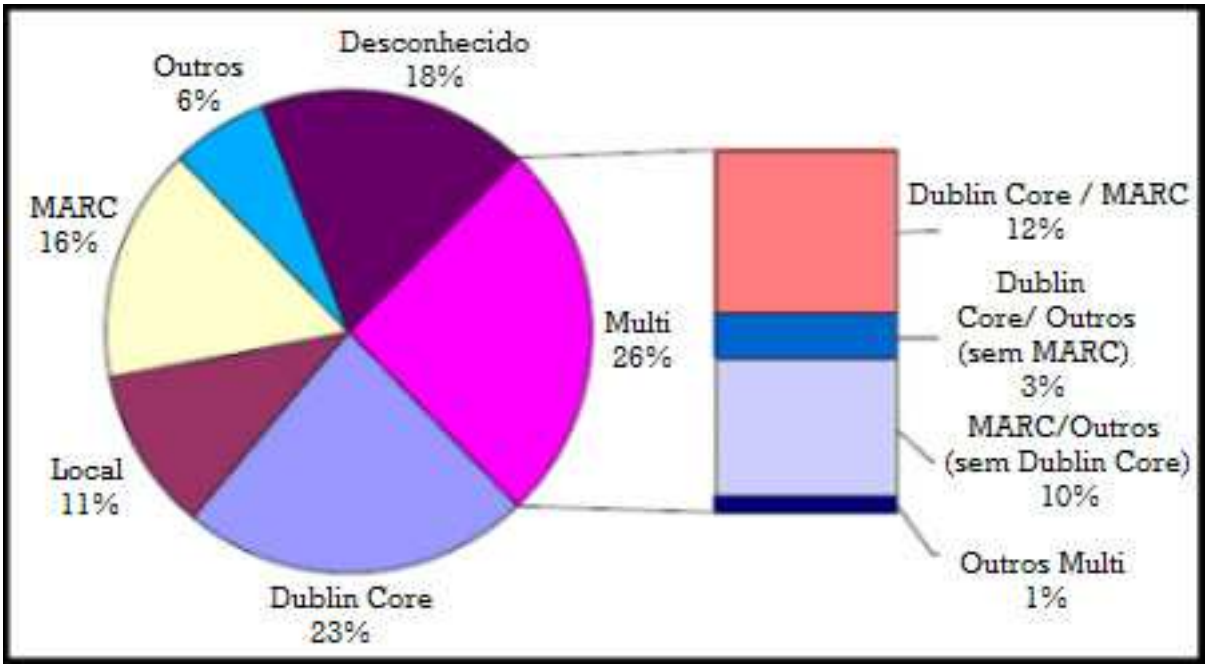

Figura 2.2 Porcentagem dos padrões de metadados usados nas Instituições.

O padrão MARC (*Machine-Readable Cataloging*) foi um dos primeiros padrões para descrição de recursos de informação, visando intercâmbio de informação via rede (mais especificamente em Bibliotecas Tradicionais). Um recurso descrito no padrão MARC é especificado em campos identificados com três dígitos numéricos, por exemplo:

- **Código 020:** ISBN
- **Código 100:** autor
- **Código 245:** título
- **Código 260:** publicação

Ao todo, o MARC possui 800 campos com códigos associados, o que torna a descrição de um recurso um trabalho muitas vezes desgastante e impraticável para grandes repositórios de informação. Para minimizar essas dificuldades, padrões mais simples de metadados têm sido desenvolvidos, em especial o padrão Dublin Core. Como será visto na próxima seção, a proposta deste tipo de padrão é definir um conjunto de metadados simples e intuitivo capaz de descrever diferentes documentos digitais. Devido a estas vantagens, o padrão de metadados adotado para os nossos estudos foi o Dublin Core.

2.2.1 Padrão Dublin Core

Em 1995, na cidade de Dublin, Irlanda, foi criado o *Dublin Core Metadata Initiative* (DCMI) com o intuito de promover a interoperabilidade entre metadados ao redor do mundo. O seu principal resultado foi a criação do padrão de metadados Dublin Core [Men05]. O Dublin Core é um padrão internacional de metadados compostos por quinze elementos básicos usados para descrever uma variedade de fontes digitais. A semântica desses elementos foi estabelecida através do consenso de vários grupos interdisciplinares, com bibliotecários, museus, editoras e analistas de sistema [Gro05].

Este padrão inclui dois níveis para podermos descrever os recursos em uma rede: o simples e o qualificado:

- **Simples:** Neste nível, utilizamos os quinze elementos básicos para descrever um recurso. Os mesmos serão vistos com detalhes ao longo desta subseção.
- **Qualificadores:** Além dos quinze elementos básicos, provenientes do nível simples do Dublin Core, este nível inclui ainda três grupos de elementos (*Audience*, *Provenance* e *RightsHolder*) que também são chamados de grupos de refinamento. Como o próprio nome sugere, estes elementos refinam a semântica dos objetos eletrônicos e podem ser bastante úteis na busca de tais recursos.

Podemos destacar aqui alguns princípios seguidos para a elaboração deste padrão de metadados [Corc]:

- **Simplicidade:** O conjunto de elementos do Dublin Core foi estabelecido para ser simples e pequeno. Desta forma, a maioria dos usuários (mesmos os não especialistas) pode

descrever um recurso facilmente, provendo assim uma recuperação facilitada desses objetos eletrônicos por outros usuários.

- **Semântica Universal:** Recuperar a informação na Internet não é algo trivial. Uma de suas razões deve-se ao fato das diferenças entre terminologias e descrições de recursos. O Dublin Core foi criado para ajudar um pesquisador não especialista a achar um recurso através de elementos que são universalmente compreendidos. Por exemplo, se um repositório científico estiver estruturado de acordo com o padrão Dublin Core e quisermos encontrar o autor de um artigo, basta procurarmos o elemento creator. Esta representação mais genérica aumenta a visibilidade e acessibilidade ao recurso eletrônico.

- **Extensibilidade:** Para algumas poucas aplicações, os conjuntos dos elementos Dublin Core não são suficientes para descrever um recurso. É esperando, então, que outras comunidades especializadas em metadados criem elementos adicionais para estes conjuntos. Tais elementos de refinamento podem ser usados juntos com os quinze elementos básicos do Dublin Core para permitir uma melhor descrição e interoperabilidade.

A tabela 2.1 apresenta os 15 elementos do padrão Dublin-Core, com sua descrição e alguns exemplos para cada elemento [Cora].

A fim de mostrarmos a vantagem da simplicidade do Dublin Core frente ao padrão MARC, observe a tabela 2.2 com a equivalência de alguns campos nos dois padrões [GYa04]. Como pode ser visto, o Dublin-Core unifica elementos que são definidos em diferentes campos do MARC.

Os elementos do Dublin-Core podem ser representados computacionalmente de diferentes maneiras (eg.: XHTML/HTML, XML, XML/RDF). A forma mais utilizada é codificarmos o recurso em formato XML [wS]. De acordo com as recomendações da [PJ03], um título de um recurso, por exemplo, pode ser expresso da seguinte maneira:

<dc:title>Dublin Core em XML</dc:title>

E poderíamos descrever os demais elementos de um recurso de forma similar ao mostrado acima. A figura 2.3 apresenta o livro *Modern Information Retrieval*, descrito no padrão Dublin Core em XML.

Deste recurso codificado em XML mostrado na figura 2.3, são realizadas algumas observações pertinentes:

- A primeira linha do código corresponde ao cabeçalho (também chamado de declaração) do documento XML, que tem a função de definir a versão do XML e a codificação do caractere utilizado no documento. No exemplo da figura 2.3, o documento está em conformidade com a versão 1.0 e usa a ISO-8859-1 (*Latin-1/West European*) como conjunto de caracteres.

- A segunda linha representa o elemento raiz do documento XML, que no caso é denominado de metadata. Os parâmetros desse elemento raiz definem os esquemas em XML, usados na construção do documento.

- O restante do documento são os metadados que descrevem o recurso exemplificado.

Tabela 2.1 Os Elementos Dublin Core

Elemento	**Descrição**	**Exemplo**
title	O nome dado ao recurso. Efetivamente, é o nome que o recurso é formalmente conhecido.	Casa-Grande e Senzala
subject	O tópico do conteúdo do recurso. Tipicamente, o subject pode ser expresso nas palavras-chave do recurso.	escravidão, literatura, senhor de engenho
description	Uma descrição do conteúdo do recurso. Esse elemento pode incluir: um resumo, um índice ou um texto livre sobre o conteúdo do recurso.	Em 1933, Gilberto Freyre publica Casa-Grande e Senzala, um livro que revoluciona os estudos no Brasil, tanto pela novidade dos conceitos quanto pela qualidade literária
type	A natureza ou o gênero do conteúdo do recurso.	texto
source	Uma referência de onde o recurso foi gerado.	Casa-Grande e Senzala, Primeira Edição
relation	Uma referência para um outro recurso relacionado.	Nordeste: Aspectos da Influência da Cana Sobre a Vida e Paisagem
coverage	Inclui um local (nome de um lugar ou coordenadas geográficas), um período no tempo ou jurisdição (entidade administrativa).	Recife-PE
creator	Entidade responsável pela criação do conteúdo do recurso. Pode ser uma pessoa, uma organização ou um serviço.	Gilberto Freyre
publisher	Entidade responsável por disponibilizar o recurso criado.	Editora Global
contribuitor	Pessoa, organização ou serviço responsável por contribuir com o conteúdo do recurso criado.	Roger Bastide
rights	Informação sobre os direitos do recurso.	Proibida Reprodução
date	Uma data que pode ser associada com a criação ou disponibilização do recurso.	1933
format	A física ou digital composição do recurso. Pode incluir um tipo de mídia ou dimensões do recurso (tamanho ou duração).	1.5 MB
identifier	Uma referência única para o recurso dado. Exemplos de identificação podem ser o ISBN de um livro, o ASIN de uma música, uma URL com o recurso etc.	ISBN: 852601059X
language	O idioma do conteúdo intelectual do recurso. É recomendável que se utilize os valores dos elementos definidos por [Gro]	pt-br

Tabela 2.2 Exemplos de equivalência entre elementos do Dublin Core e campos do MARC

Elementos Dublin Core	**Campos Marc**
Title	245ab
Creator	100,110,111, 710, 711
Subject	600, 610, 611, 630, 650, 653
Publisher	260$b
Date	260$c

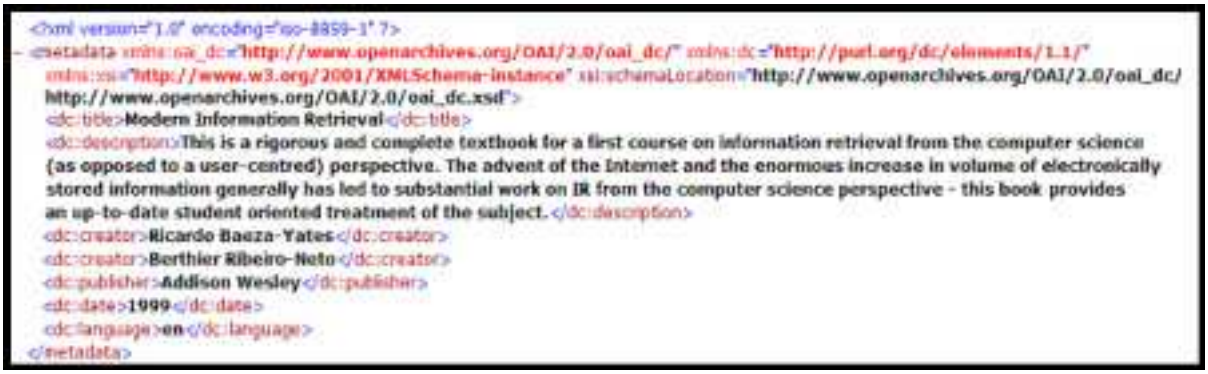

Figura 2.3 Um recurso codificado em XML.

2.3 Integração de Dados

A informação é o bem mais precioso que uma organização pode possuir [Sak05]. Desta forma, disseminar informação de maneira integrada e organizada é multiplicar a riqueza de qualquer instituição. Recentemente, temos observado um crescente interesse na área de integração de dados, em absolutamente todas as áreas de conhecimento. A integração de informações oferece um meio menos dispendioso e mais flexível de acessar dados a partir de fontes diversas. Como exemplo, podemos citar algumas áreas que utilizam a integração de dados em larga escala:

- **Business Intelligence:** Com uma plataforma de dados integrados em uma empresa, é possível desenvolver um ambiente de Business Intelligence com base em informações confiáveis, e tornar o gerenciamento das mudanças que ocorrem na corporação uma tarefa mais fácil e ágil, facilitando o retorno sobre o investimento [Obj].

- **Saúde:** A necessidade de se conhecer informações sobre a saúde de um contingente populacional é mais do que evidente. A análise adequada e a disponibilização oportuna das informações em saúde subsidiaria os processos de acompanhamento e avaliação das condições de saúde de um conjunto de beneficiários, das ações e serviços que lhes são prestados. As informações coletadas e integradas são de fundamental importância na geração de indicadores da situação de saúde da população, fornecendo subsídios para que o governo conheça e promova melhorias no setor [Lou03].

- **Repositórios Científicos:** Atualmente, diversas organizações ao redor do mundo estão criando repositórios para material científico, como teses, dissertações, artigos e monografias. Desta forma, pretende-se que estudantes e pesquisadores tenham acesso facilitado e integrado a diversas fontes acadêmicas. Um dos exemplos que podemos citar é o projeto Biblioteca Digital de Teses e Dissertações [IBI]. O projeto busca integrar sistemas de informação existentes nas Instituições de Ensino Superior brasileiras, bem como estimular o registro e a publicação de teses e dissertações em meio eletrônico.

Esses três exemplos é apenas uma amostra da diversidade de áreas em que a integração de dados está sendo estudada e aplicada.

O principal objetivo de um sistema integrador de dados é disponibilizar ao usuário uma interface única para o acesso a informações disponíveis em diferentes bases de dados. Desta forma, é especificado o que se deseja procurar e o sistema determina onde a informação pode ser encontrada, retornando ao usuário as respostas de acordo com a sua consulta [SL01].

Uma característica importante dos sistemas de integração de dados é que eles trabalham com fontes de dados heterogêneas e que podem ser autônomas, conforma mostra a figura 2.4.

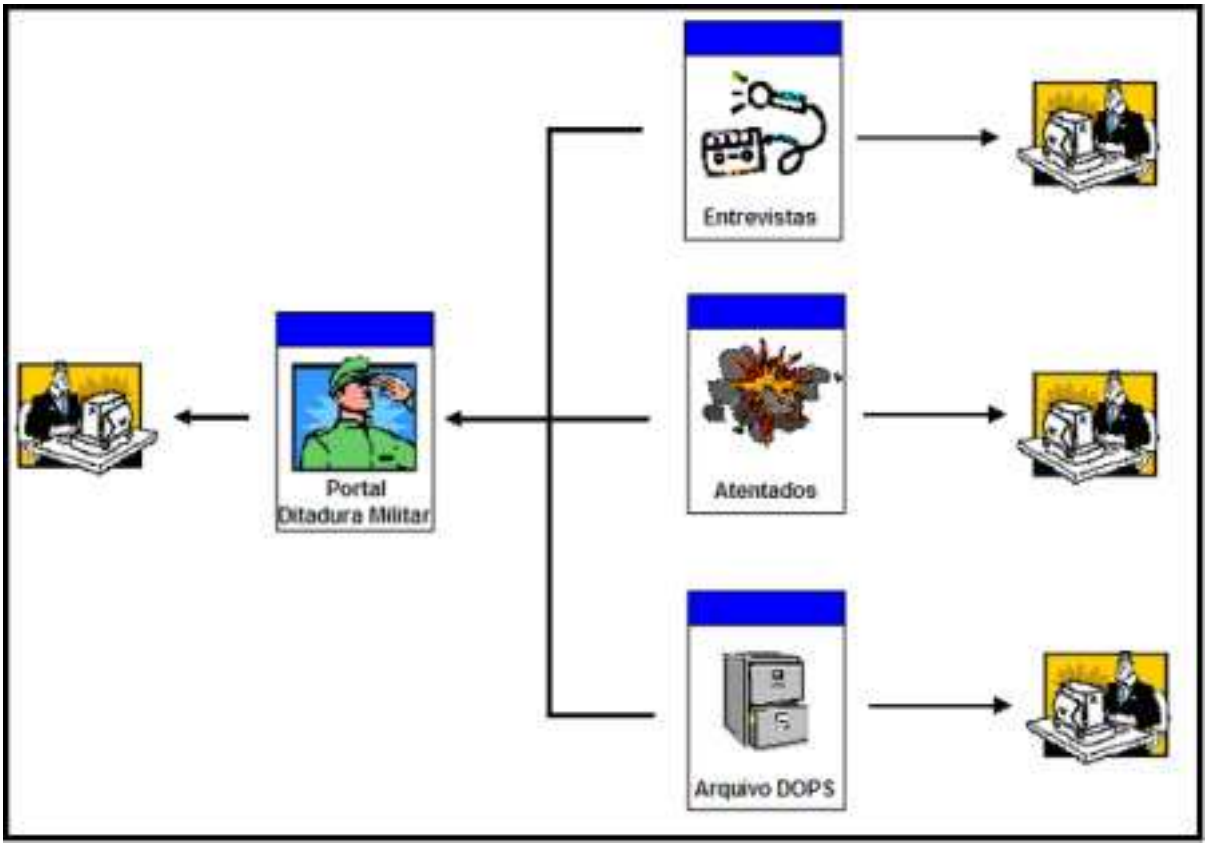

Figura 2.4 Exemplo de um sistema integrador de dados.

No exemplo mostrado, existem três Bibliotecas Digitais com temas específicos sobre a ditadura militar brasileira (entrevistas com militantes, arquivos sobre atentados e arquivos do DOPS, a delegacia da época). Esses repositórios devem suportar aplicações locais, permitindo que usuários pesquisem em uma Biblioteca Digital específica. Ao mesmo tempo, devem disponibilizar informação para um sistema integrador, no caso, um portal sobre a ditadura militar brasileira. Assim, um usuário teria a opção de pesquisar arquivos destas três fontes através de um único acesso.

Nas próximas seções, veremos alguns conceitos, abordagens e arquiteturas para o melhor tratamento da integração de dados.

2.3.1 Abordagens e Arquiteturas para Integração

Muitas arquiteturas foram propostas para minimizar o problema da complexidade dos sistemas para integração e interoperabilidade de dados. Citaremos nesta subseção algumas arquiteturas típicas que foram propostas com esta finalidade. As arquiteturas para este fim são baseadas numa abordagem multicamadas, como pode ser observado na figura 2.5 [ABS00].

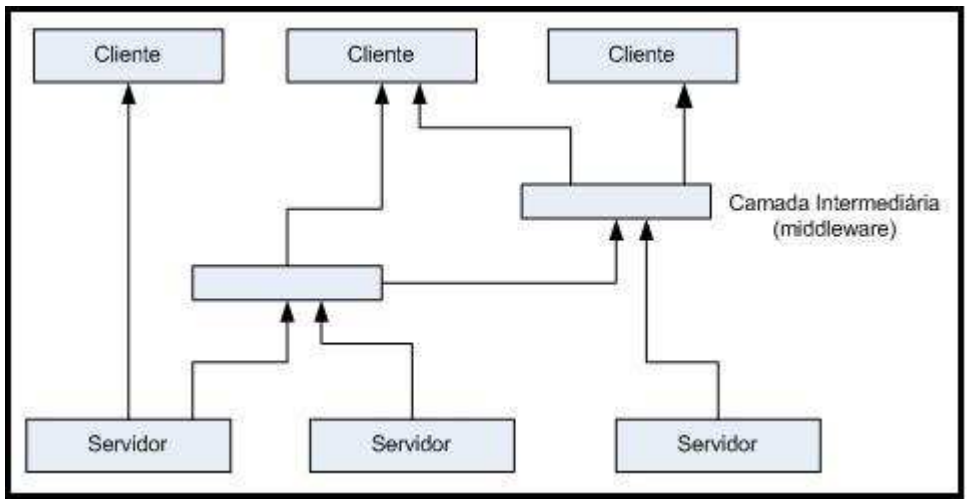

Figura 2.5 Uma arquitetura multicamadas.

Nas camadas de mais baixo nível, estão os chamados servidores, que são as fontes de dados, podendo ser servidores de banco de dados ou de arquivos, sistemas de informações ou qualquer aplicação que produz dados (uma Biblioteca Digital, por exemplo). A camada superior consiste na camada do cliente, contando com as interfaces ou sistemas que consomem os dados dos servidores. Entre as duas camadas, encontra-se uma camada (ou conjunto de camadas) intermediária chamada de *middleware*, que é um software para integrar esses dados. A partir desta arquitetura geral, apresentado na figura 2.5, podemos encontrar diversas outras arquiteturas com o foco voltado para a integração de dados.

Para a construção de qualquer sistema para integração de dados, uma decisão crucial é a escolha da abordagem utilizada: virtual ou materializada [SL01]. E cada uma das duas abordagens citadas possui uma arquitetura específica. As arquiteturas utilizadas nas abordagens virtuais e materializadas são, respectivamente, as arquiteturas de Mediadores e a de Data Warehouse.

Veremos a seguir uma breve discussão sobre as duas abordagens citadas e suas arquiteturas.

Abordagem Virtual

Neste tipo de abordagem, as fontes que contêm os dados só são requisitadas quando as consultas são efetuadas. Uma grande vantagem deste tipo de abordagem é que as informações estão sempre atualizadas, já que a resposta da consulta é requisitada à fonte original. Por outro lado, esta abordagem não é recomendada quando os servidores forem instáveis, pois pode deixar os

serviços inacessíveis [SL01]. Assim, concluímos que devemos usar essa abordagem quando as informações mudam rapidamente e for de vital importância termos esses dados sempre atualizados. Como exemplo para este tipo de abordagem, poderíamos citar os sistemas bancários. Nesses sistemas, há diversas transações em um espaço de tempo muito curto e tais informações precisam estar sempre atualizadas para o cliente.

Esta abordagem utiliza a arquitetura de mediadores, conforme dito anteriormente. Na arquitetura de mediadores [SL01], o acesso aos dados distribuídos em múltiplas fontes é efetuado através de consultas que são submetidas ao sistema via um mediador. Este mediador transforma cada consulta em subconsultas a serem enviadas às fontes de dados. As subconsultas geradas pelo mediador devem ser traduzidas para a linguagem de consulta de cada fonte de dados a ser consultada. Ao final, os resultados das subconsultas são integrados e a resposta é devolvida ao usuário (Figura 2.6) [ABS00].

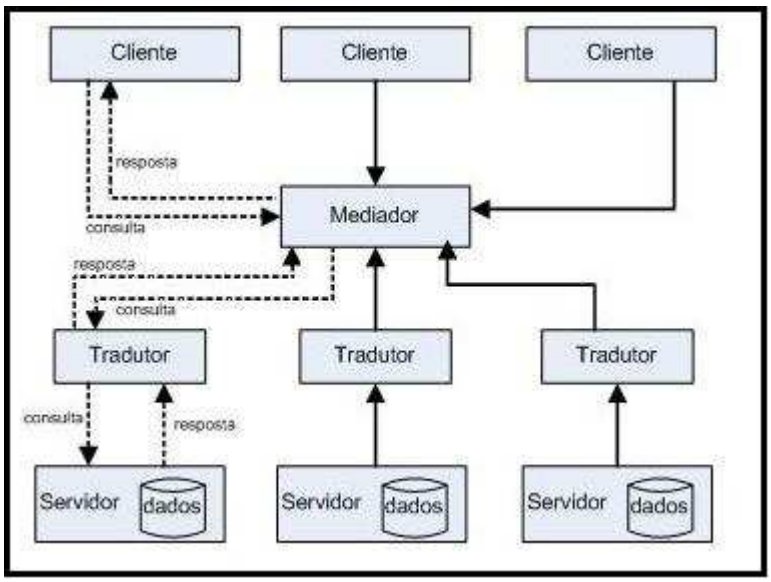

Figura 2.6 Arquitetura de Mediadores.

Outros componentes desta arquitetura são os tradutores, que convertem tanto os dados das fontes para um modelo de dados comum, como também consultas de aplicações em consultas específicas da fonte de dados correspondente.

Abordagem Materializada

Diferente da primeira, a abordagem materializada recupera as informações das fontes de dados previamente, e as armazena em um repositório central. O usuário faz, então, sua consulta a esse repositório. A grande vantagem deste tipo de abordagem é que as informações estão sempre disponíveis para a consulta, tornado o sistema mais eficiente e com um melhor desempenho. Por outro lado, a sua principal desvantagem é que os dados não estão sempre atualizados. Assim, esta abordagem não é adequada quando as informações integradas precisam estar sempre atualizadas. Esta abordagem é mais adequada quando: são requisitadas porções específicas e

previsíveis da informação disponível; usuários demandam altos desempenho de consulta, sem requerer que o estado da informação seja o mais recente (e.g., engenhos de busca na Web); usuários querem guardar informações que não são mantidas nas fontes de dados originais, tais como informações históricas.

Neste tipo de abordagem, utilizamos a arquitetura de Data Warehouse, como mostra a figura 2.7.

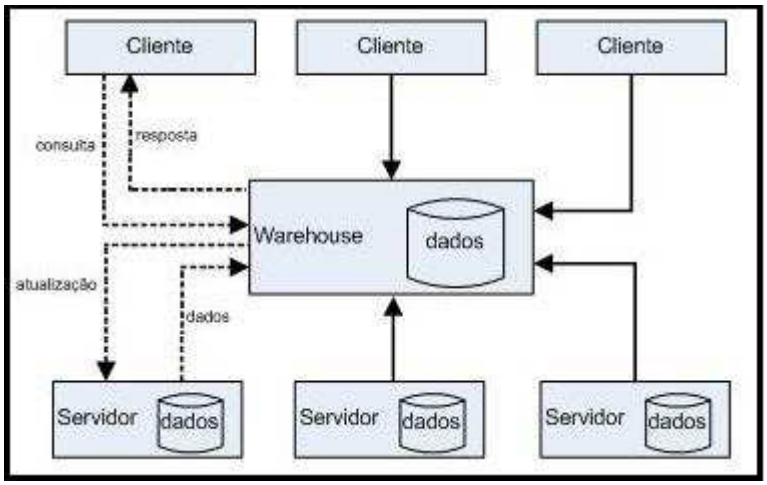

Figura 2.7 Arquitetura de Data Warehouse.

Como informado anteriormente, para um sistema integrador de dados, as suas fontes também exercem aplicações locais e o seu repositório certamente sofrerá modificações. Desta maneira, encontramos duas opções para a manutenção de um sistema com arquitetura Data Warehouse [SL01]:

- **Rematerialização da visão:** O conteúdo inteiro do data warehouse é apagado e toda a informação (tanto novas quanto antigas) das fontes de dados são carregados novamente no Data Warehouse.
- **Manutenção incremental:** Mudanças nos dados das fontes de dados locais são propagadas incrementalmente (à medida que os dados são inseridos ou alterados) para o Data Warehouse.

Encontramos na tabela 2.3 um resumo das duas abordagens apresentadas, apresentado as vantagens e desvantagens, bem como em que situação os seus usos são mais adequados.

Tabela 2.3 Abordagem Virtual X Abordagem Materializada

	Abordagem Virtual	**Abordagem Materializada**
Vantagens	Informações recuperadas sempre atualizadas	Informações disponíveis imediatas para consulta
Desvantagens	Fontes de dados podem estar temporariamente inacessíveis	Consistência deve ser mantida entre o repositório e a fonte
Adequada	Quando informações mudam rapidamente	Usuários demandam alto desempenho
Inadequada	Se fontes podem ficar fora do ar por alguma razão	Se as informações precisam estar sempre atualizadas

2.3.2 Integração de Bibliotecas Digitais

A ausência de integração entre Bibliotecas Digitais é considerada um dos problemas mais significativos encontrados hoje em dia neste tipo de sistema, sendo um dos temas mais comentados e trabalhados pela comunidade especializada [LMZN01].

O crescimento da necessidade de recursos que, não apenas possuem metadados descritivos, mas que expõe seus dados para a coleta vem crescendo de acordo com o avanço tecnológico. Há dois casos principais que motivam essa necessidade [dSNLW04]:

- **Preservação digital:** A transferência de recursos digitais de uma fonte de dados para uma outra, armazena cópias desses dados, ajudando na preservação do mesmo.
- **Descoberta facilitada:** Se uma Biblioteca Digital agrega fontes diversas, um usuário consegue encontrar mais fácil o recurso pesquisado.Ou seja, como o objeto procurado pelo pesquisador também se encontra na fonte de dados original, resta mais de um local para pesquisa.

Para a eficiente integração de Bibliotecas Digitais, deve-se planejar uma arquitetura robusta, definir um padrão de metadados apropriado e um conjunto de protocolos para comunicação. Tais requisitos servirão para prover uma interface uniforme, escondendo especificações e restrições das fontes de dados, fornecendo uma visão integrada dos mesmos [PPG02].

Diversos protocolos já foram criados para a integração de Bibliotecas Digitais, podendo-se destacar alguns:

- **Z39.50:** O Z39.50 é um padrão para recuperação da informação, que especifica as estruturas de dados e regras de interconexão de recursos. Foi desenhado para permitir pesquisa e recuperação de informação em redes de computadores distribuídos. É baseado na arquitetura cliente/servidor e opera sobre a rede Internet [Ros97]. O Z39.50 utiliza a abordagem virtual para integração de dados.
- **SDLIP:** Desenvolvida por um grupo de universidades americanas, o protocolo SDLIP (Simple Digital Library Interoperability Protocol) foi criado com o intuito de definir um padrão mais simples que o Z39.50 [PBJ+00]. Possui algumas características como simplicidade na implementação do cliente e do servidor dos dados e possibilidade de implementação com tecnologias para objetos distribuídos (e.g., CORBA).

- **NCSTRL:** O NCSTRL não é apenas um protocolo para integração de dados. É uma organização com mais de 100 instituições com o objetivo de prover um conjunto de bibliotecas integradas. Entretanto, é voltada para matérias sobre Ciência da Computação. [Leiner, 1998].

- **OAI:** O *Open Archives Initiative* (OAI) tem o objetivo de desenvolver e promover padrões de interoperabilidade que visam facilitar a disseminação eficiente de conteúdo. Tal movimento teve início com o objetivo de ampliar o acesso a bases de dados de artigos científicos. Os principais padrões e ferramentas desenvolvidas, contudo, não dependem do tipo de conteúdo que é oferecido. O seu protocolo, o OAI-PMH, é de fácil implementação e baseado em padrões já existentes (XML, HTTP e Dublin Core) [Car05]. Vale ressaltar que o OAI-PMH utiliza a abordagem materializada para realizar a integração entre os seus repositórios.

Nos dias de hoje, os protocolos mais utilizados para a interoperabilidade entre Bibliotecas Digitais são o Z39.50 e o OAI-PMH. A diferença crucial entre eles é a facilidade no uso do protocolo. Enquanto o Z39.50 trabalha com uma abordagem virtual e possui inúmeras regras para o seu pleno funcionamento, o OAI operacionaliza de acordo com a abordagem materializada, o que facilita a sua implementação e entendimento. De fato, para sistemas de Bibliotecas Digitais, as vantagens da abordagem materializada (eficiência na consulta e facilidade de implementação) sobre a virtual (bases para consulta sempre atualizadas) justificam o uso da primeira, facilitando a descoberta de dados em fontes diversas.

Segundo [LMZN01], o *Open Archives* difere de outras abordagens para interoperabilidade entre Bibliotecas Digitais por ser a mais simples, completa e com protocolo de fácil implementação. O framework do OAI define dois papéis fundamentais: os Provedores de Dados (expõe os recursos em um formato específico) e os Provedores de Serviços (coletam a informação dos provedores de dados através do protocolo OAI-PMH).

Desta maneira, percebendo as vantagens do OAI em nosso tema de pesquisa, esta iniciativa foi a escolhida para o nosso trabalho. E, pela sua fundamental importância, o Open Archives Initiative será visto com detalhes em um capítulo à parte, o capítulo 3.

2.4 Critérios de Avaliação

A partir da segunda metade da década de 1990, as pesquisas no campo das Bibliotecas Digitais deram grande crescimento [WMBB00]. Estes sistemas, entretanto, não eram bem avaliados, pois usualmente alguns consideravam que a função de uma Biblioteca Digital era a simples disponibilização da informação na Web.

Com isso, pretendemos aqui definir critérios relevantes para avaliar se uma Biblioteca Digital atende bem a seus usuários. Para alçarmos esse objetivo, analisamos diferentes Bibliotecas Digitais disponíveis na Internet, para formularmos com mais embasamento tais critérios.

A Biblioteca Digital da figura 2.8 trata-se do Museu História Nacional[1]. Observe no detalhe do canto superior direito uma consulta pouco convencional: utilizando o artifício do navegador,

[1]Museu Histórico Nacional - http://www.museuhistoriconacional.com.br

o CTRL+F, para proceder a "busca".

Figura 2.8 Biblioteca do Museu Histórico Nacional.

No segundo exemplo encontrado na figura 2.9 é apresentado a Biblioteca Nacional Digital de Portugal[2] . Na figura 2.9, encontramos um sistema de busca não muito trivial, com diversos campos a serem preenchidos. Além disso esta Biblioteca Digital apenas fornece a referência do documento e a ida a uma biblioteca tradicional torna-se muitas vezes indispensável neste caso.

[2]Biblioteca Nacional Digital de Portugal - http://pacweb.bn.pt/bnd.htm

Figura 2.9 Biblioteca Nacional de Portugal.

Um outro exemplo de Biblioteca Digital interessante é a Biblioteca Virtual do Rio Grande do Sul[3], mostrada na figura 2.10. Aqui existe a possibilidade da busca pelo conteúdo da informação, porém tal sistema é muito complexo para usuários leigos, que desconhecem álgebra *Booleana*.

Figura 2.10 Biblioteca Virtual do Rio Grande do Sul.

[3]Biblioteca Virtual do Rio Grande do Sul - http://www.bibvirtual.rs.gov.br

Uma das maiores Bibliotecas Digitais para documentos históricos do mundo, a *Library of Congress* (LOC)[4] possui um projeto chamado *American Memory*[5]. A biblioteca possui um sistema de busca e os resultados são retornados ao usuário rapidamente. Contudo, o sistema de visualização do documento é feito de forma bastante simples, como mostrado na figura 2.11. Mesmo assim, essa Biblioteca Digital é tomada como referência, pois, como dito anteriormente, dificilmente os sistemas disponibilizam o documento para a visualização.

Figura 2.11 Library of Congress.

[4]Library of Congress - http://www.loc.gov/
[5]America Memory from the Library of Congress - http://memory.loc.gov/ammem/

Por fim, apresentamos na figura 2.12 o *CiteSeer*[6], uma Biblioteca Digital de literatura científica. A sua característica principal é que as informações nela contida são coletadas de diversos repositórios, o que o torna um sistema integrador de dados. Um sistema como este, entretanto, poderia conter informações científicas em diversos tipos de mídias (e.g.: aulas em vídeo, entrevistas em áudio). O *CiteSeer*, entretanto, trabalha apenas com informações textuais.

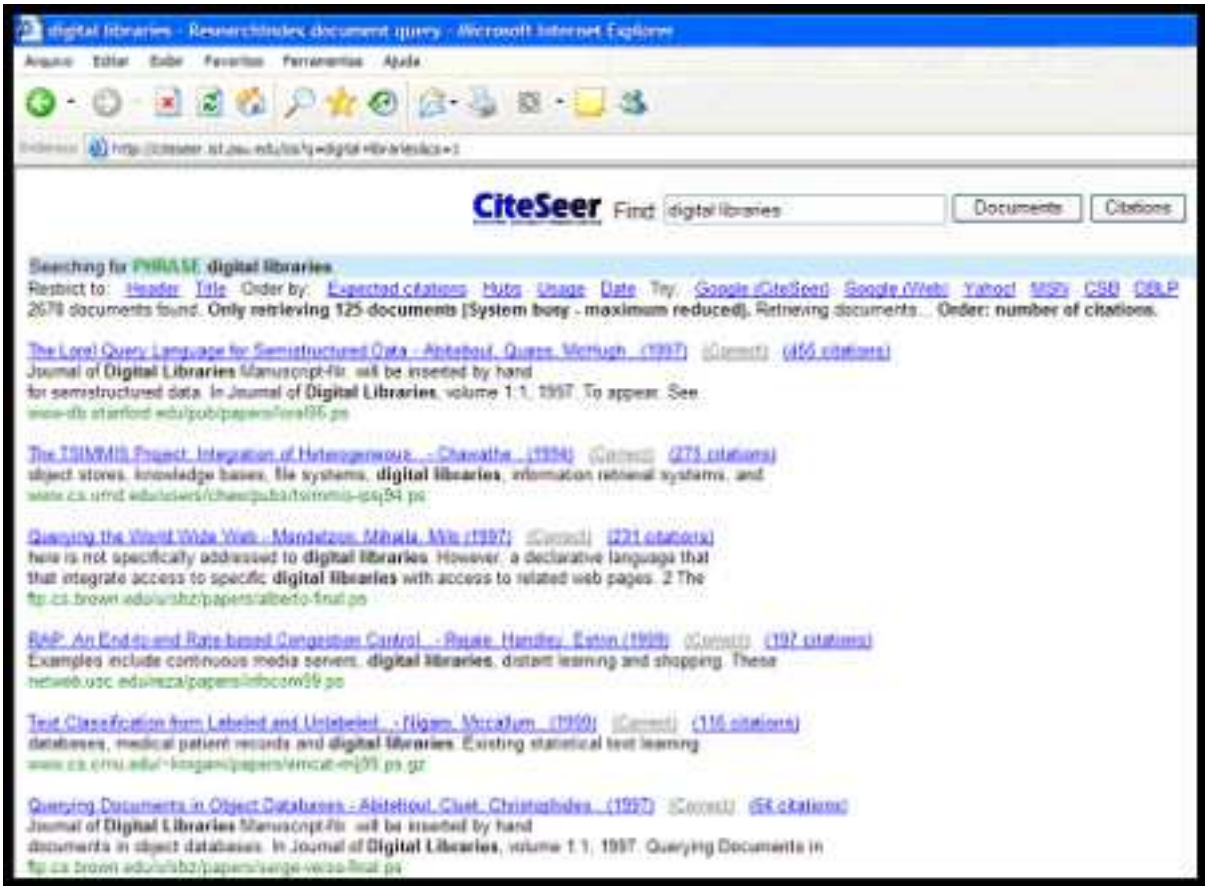

Figura 2.12 Biblioteca Digital do CiteSeer.

A partir da análise realiza nessas Bibliotecas Digitais, definimos abaixo alguns critérios de avaliação para este tipo de sistema:

- **Sistema de Busca:** Como já dito anteriormente, a função de uma Biblioteca Digital não é simplesmente disponibilizar informação na Internet. Devemos prover ao usuário algum mecanismo de busca para que o mesmo possa encontrar com maior facilidade o que necessita. E, para que a consulta seja satisfatória para o pesquisador, é recomendável também que esse sistema de busca retorne os dados ao usuário por ordem de relevância. Além do mais, essa consulta deve ser simples (e.g. um campo de busca para todos os

[6]CiteSeer - http://citeseer.ist.psu.edu/

metadados), oferecendo a opção de uma busca mais refinada para usuários mais avançados.

- **Visualização dos Documentos:** É de fundamental importância que o usuário que visita uma Biblioteca Digital encontre o que está procurando, ou seja, a informação completa. Dessa maneira, é recomendável que a Biblioteca Digital possua o documento completo para a total satisfação do usuário. Naturalmente, quando estamos lendo um livro e as suas letras estão pequenas, recorremos à ajuda de uma lente de aumento. Se em alguma página do mesmo livro uma imagem estiver invertida em relação a sua posição normal, giramos o livro na direção que nos convir. E desta forma deveria ser uma Biblioteca Digital. Assim, é preferível que os usuários sejam capazes de manipular os documentos (e.g.: zoom, inverter, clarear, etc.) para uma melhor visualização da informação.
- **Biblioteca Multimídia:** Com o avanço da tecnologia, podemos representar a informação em diversos formatos (e.g. áudio, vídeo, texto ou imagem). Desta maneira, uma Biblioteca Digital também deve ser capaz de oferecer ao usuário essa variedade de tipos de informação.
- **Integração de Dados:** Bibliotecas Digitais que compartilham os seus acervos oferecem um importante avanço para usuários que desejem acessar, em um só local, diversas bases de domínios específicos. Este é um dos principais critérios de avaliação adotado, visto a sua importância na disseminação da informação pela Internet.

Mostrado os critérios de avaliação adotados para as Bibliotecas Digitais, relacionamos cada Biblioteca Digital apresentada anteriormente com os seus critérios de avaliação na tabela 2.4.

Tabela 2.4 Relação das Bibliotecas Digitais com os critérios de avaliação

	Sistema de Busca	**Visualização dos documentos**	**Biblioteca Multimídia**	**Integração de Dados**
Biblioteca do Museu Histórico Nacional	Não	Não	Não	Não
Biblioteca Nacional de Portugal	Sim	Não	Não	Não
Biblioteca Virtual do RS	Sim	Sim	Não	Não
Library of Congress	Sim	Sim	Não	Não
CiteSeer	Sim	Sim	Não	Sim

2.5 Considerações Finais

Bibliotecas Digitais são mais do que sistemas de depósito de documentos digitalizados. Hoje em dia, o tema está sendo muito discutido entre os especialistas da área e definições são con-

tinuamente atualizadas. Apesar do esforço, muitas são as barreiras que um usuário enfrenta ao pesquisar um documento em uma Biblioteca Digital: ausência ou ineficiência de um sistema de buscas, documento não disponibilizado por completo e a não aceitação de diferentes tipos de mídia como áudio e vídeo.

Além disso, é de grande necessidade que as Bibliotecas Digitais estruturem de maneira correta as suas informações para a eficiente disseminação do conteúdo ao alcance de todos. Uma parte desse problema pode ser resolvido com o uso de metadados para descrever os recursos eletrônicos de um repositório.

O uso de metadados também é essencial para um dos assuntos mais pesquisados na área de Bibliotecas Digitais atualmente: a interoperabilidade entre estes tipos de sistemas. Realizando este requisito, o usuário terá um ponto único de acesso a diversas fontes diferentes, facilitando o encontro da informação desejada. Adicionalmente, a interoperabilidade entre bases colabora com a virtual preservação do documento eletrônico e amplia as formas de encontrá-lo.

Para atingir a interoperabilidade entre Bibliotecas Digitais, muitos protocolos foram criados. Um deles, entretanto, destaca-se pela facilidade de implementação e completude de suas funcionalidades: o OAI-PMH, do *Open Archives Initiative*. Apesar de ainda novo, muitas Bibliotecas Digitais já estão implementando o protocolo e os resultados são satisfatórios. Entretanto, há ainda muito por fazer nesta área e, pela importância do assunto, detalhes do *Open Archives Initiative* e suas aplicações serão vistos com detalhes no capítulo seguinte, além de formularmos novos critérios para avaliação de Bibliotecas Digitais.

CAPÍTULO 3

Open Archives Initiative - OAI

As teorias científicas e artísticas contemporâneas não pensam mais a realidade em grupos de diferentes objetos, separados de nós, mas em grupos de diferentes integrações que incluem o observador.

—ANDRÉ PARENTE

Com a visão voltada para a eficiente disseminação de informações através da Internet,a *Open Archives Initiative* (OAI) promove e desenvolve padrões para a interoperabilidade entre repositórios digitais [Brab]. O uso do OAI tornou-se um importante aliado para a integração entre Bibliotecas Digitais, dado a simplicidade e eficiência de seu protocolo, o *Open Archives Initiative Protocol for Metadata Harvesting* (OAI-PMH).

O objetivo deste capítulo é apresentar os aspectos mais relevantes sobre esta iniciativa, contemplando os componentes do OAI (Provedores de Dados e Provedores de Serviços) e o protocolo OAI-PMH. Ainda realizaremos um estudo a respeito de algumas ferramentas e Bibliotecas Digitais que utilizam o Open Archives e definiremos alguns critérios que acreditamos serem relevantes para a implementação de Bibliotecas Digitais integradas.

3.1 Histórico e Descrição Geral

No final dos anos 90, já existiam alguns repositórios na Internet em diversas áreas, principalmente artigos científicos. Apesar da utilidade e importância desses repositórios, dois aspectos dificultam a disseminação dos documentos armazenados: (1) o usuário encontrava diferentes interfaces, tornando o processo de busca mais difícil; e (2) não havia uma forma automática de compartilhar os dados. Assim, Paul Ginsparg, Rick Luce e Herbert Van de Sompel, os três do Laboratório de Los Alamaos, realizaram um encontro com especialistas em outubro de 1999 em Santa Fé, Novo México. Esse encontro teve a finalidade de discutir problemas e encontrar soluções para a questão da integração entre bases na Web [LdS01]. O resultado desta conferência foi a formação do OAI (*Open Archives Initiative*), visando a criação de padrões e estruturas para permitir o acesso a uma grande quantidade de documentos eletrônicos arquivados na Internet [Gar03].

Na expressão *Open Archives*, o termo *Archive* é usado no sentido de repositórios de recursos de informação, em geral, documentos de textos, imagens, vídeos, dentre outros. Por sua vez, o termo *Open* não significa acesso gratuito e ilimitado às informações presentes nos repositórios, mas aberto em termos da arquitetura proposta pelo OAI [LdS01]. Em outras palavras, o protocolo da Open Archives é que é aberto.

O movimento OAI teve as suas raízes no esforço de ampliar o acesso a repositórios de artigos científicos como um meio de aumentar a disponibilidade da comunicação científica. Hoje, os padrões e ferramentas tecnológicas desenvolvidas, entretanto, independem do conteúdo que é disponibilizado, não se restringindo apenas a repositórios científicos [OAIb]. No registro oficial da OAI [OAIc], há também repositórios sobre biologia, arquivos históricos e vídeos, dentre outros temas, como mostra a tabela 3.1.

Tabela 3.1 Exemplo de alguns repositórios registrados na OAI

Nome do Repositório	**Descrição**	**URL**
Animal Diversity Web	Base de dados on-line da história de diversos animais, classificação e conservação da biologia da Universidade de Michigan.	http://animaldiversity.ummz.umich.edu/
Carnegie Mellon University Informedia Public Domain Video Archive	Biblioteca Digital com vídeos sobre diversos assuntos da comunidade científica.	http://infsearch.cs.cmu.edu/idvl.htm
British History Online	Biblioteca Digital contendo arquivos da época medieval e moderna britânica.	http://www.british-history.ac.uk/
Hrcak - Portal of scientific journals of Croatia	Portal da comunidade científica croata composta por artigos e jornais científicos.	http://hrcak.srce.hr/

O *Open Archives Initiative* possui algumas características principais, que são [LdS01]:

- Acesso aberto aos recursos de informação (ou pelo menos aos metadados que descrevem os recursos);
- Interface consistente entre os repositórios e os seus coletores de dados;
- Baixa barreira do protocolo, ou seja, exigência de menos esforço para a sua implementação, por se basear tecnologias já bastante difundidas (e.g.: HTTP, XML, Dublin Core).

Vale salientar que não só os metadados descritivos, mas os recursos também podem ser disponíveis para coleta, apesar de não ser uma exigência do protocolo. Por exemplo, podemos ter um repositório de vídeos que disponibiliza apenas os metadados do recurso (e.g: título do vídeo, autor, resumo) ou o vídeo em si juntamente com os seus metadados. O OAI-PMH também pode ser usado para grupos fechados, compartilhamento dos seus metadados e em aplicações comerciais.

Conforme já mencionado, a base da Iniciativa é o protocolo OAI-PMH, que faz com que os participantes da Iniciativa possam compartilhar seus metadados através de regras bastante claras e simples. Além destas regras, há dois grupos "participantes": os Provedores de Dados e os Provedores de Serviços. Os Provedores de Dados são participantes que mantêm repositórios

de informação, e implementam o protocolo OAI-PMH para expor os metadados de seus documentos. Já os Provedores de Serviço utilizam o protocolo para coletar os metadados dos Provedores de Dados, armazená-los em uma base centralizada, e oferecer algum serviço de valores agregado [LdS01].

O funcionamento básico do protocolo OAI-PMH pode ser visto na figura 3.1 [Forc]. Os Provedores de Dados armazenam os metadados de vários recursos, e são requisitados pelos Provedores de Serviços, através de requests em HTTP. Os Provedores de Serviço realizam a coleta (também chamado de *havesting*) dos metadados codificados em XML, e fornecem serviços (e.g.: unificação de metadados, busca, visualização dos documentos, etc.) para o usuário final.

Figura 3.1 Funcionamento básico do OAI-PMH.

Por se tratar de um protocolo simples, baseado em HTTP e XML, o OAI-PMH permite flexibilidade no seu desenvolvimento. Sistemas baseados no OAI podem ser desenvolvidos em uma variedade de configurações [Fora]. Como exemplo, na figura 3.2, cada Provedor de Serviços realiza a coleta em múltiplos Provedores de Dados, conforme interesses específicos dos usuários finais.

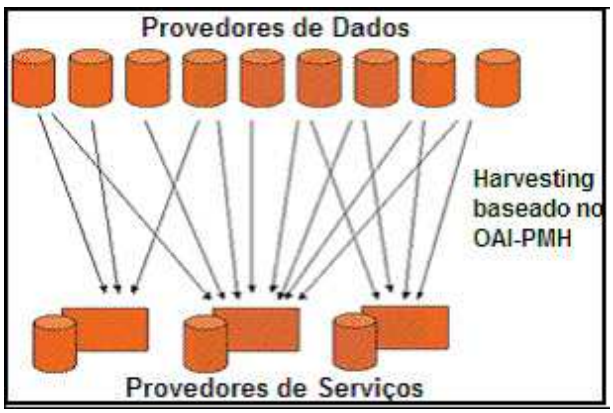

Figura 3.2 Múltiplos servidores coletando múltiplos Provedores de Dados.

Podemos também visualizar uma camada intermediária entre os Provedores de Dados e os Provedores de Serviços. Tais camadas funcionariam como agregadores de alguns repositórios específicos, que seria por sua vez colhidos por outros Provedores de Serviços (ver figura 3.3).

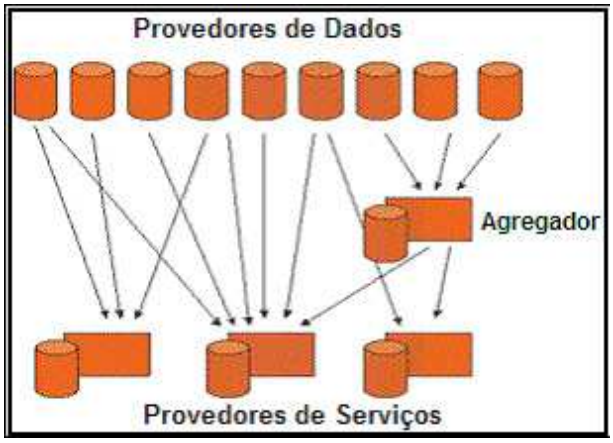

Figura 3.3 Agregador entre os provedores de dados e de serviços.

Ainda podemos ter a coleta dos dados sendo complementada com outros protocolos de acesso à informação, como o Z39.50. Este exemplo de aplicação utilizando o OAI-PMH pode ser visto na figura 3.4.

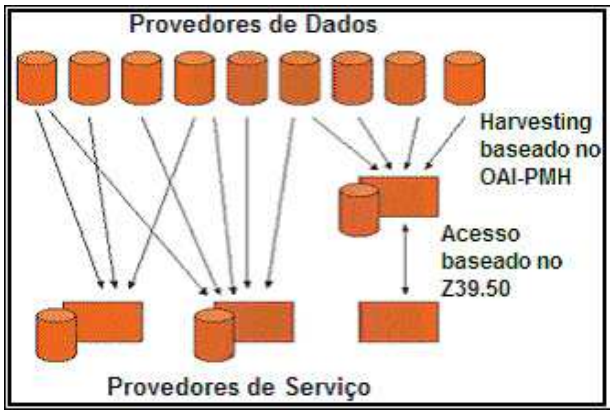

Figura 3.4 Funcionamento do OAI-PMH com a adição de outro protocolo.

3.2 Provedor de Dados

Como dito, os Provedores de Dados expõem, através do protocolo OAI-PMH, os metadados dos seus recursos, para futura coleta. O conteúdo do recurso (e.g.: texto completo, um vídeo, uma imagem) não é necessariamente exposto. Como alternativa, é incluído um link para o documento em um dos campos dos metadados [Gar03].

Desde a primeira versão do protocolo OAI-PMH, em 2001, cresce o número de repositórios registrados oficialmente no Open Archives. A figura 3.5 mostra o gráfico da ascensão do número de repositórios registrados no site oficial [1].

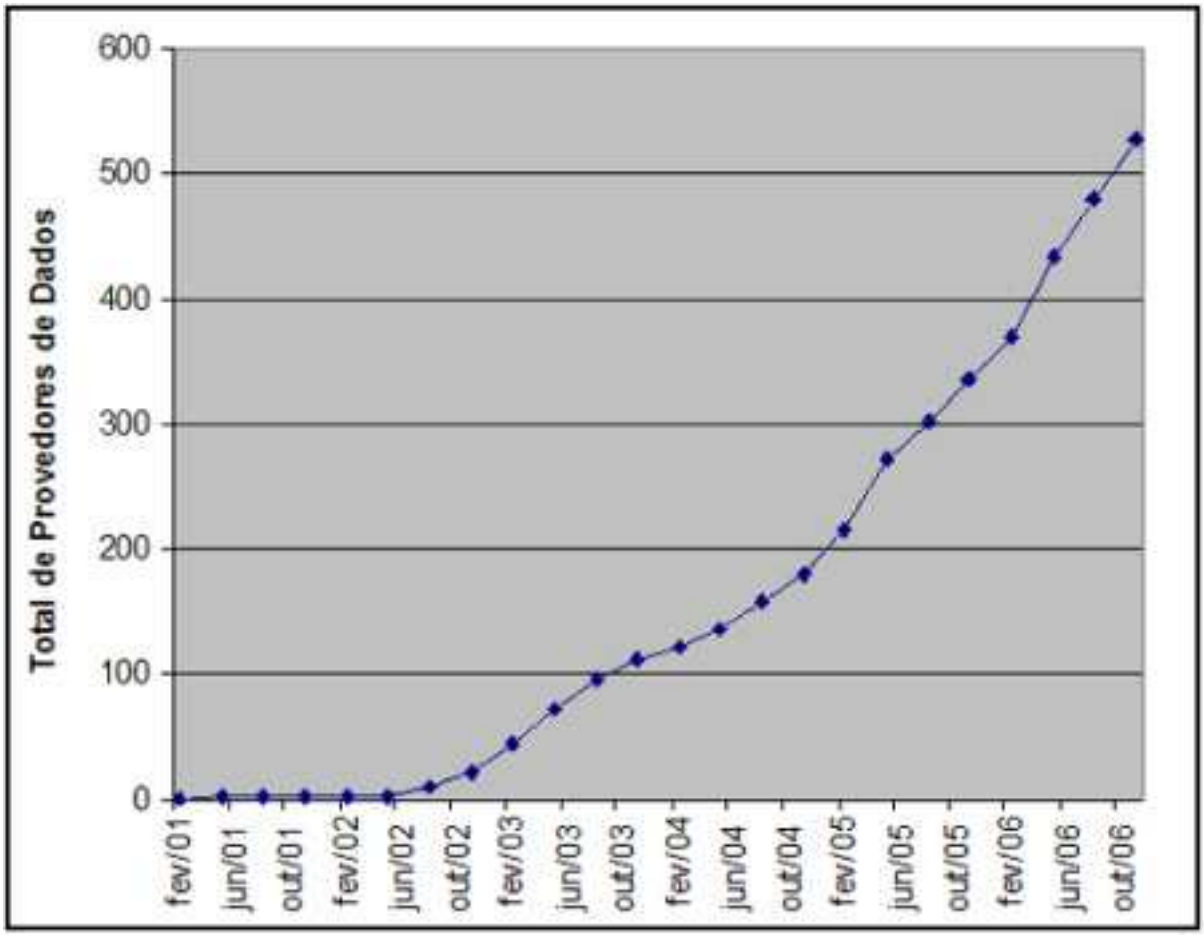

Figura 3.5 Crescimento dos provedores de dados na Web.

Como podemos observar no gráfico, o grande crescimento de repositórios trabalhando com OAI deu-se em meados no ano 2002, quando foi lançada a segunda e definitiva versão do protocolo. De fato, a primeira versão do protocolo OAI-PMH foi utilizada para testes e refinamentos de seus idealizadores [LdS01].

3.2.1 Pré-requisitos

Para implementar um Provedor de Dados de acordo com o protocolo OAI-PMH, o desenvolvedor deve estar atento a alguns conceitos e necessidades básicas que eles devem possuir para um perfeito funcionamento [Forb].

[1] Dados adquiridos por solicitação ao OAI, através do e-mail openarchives@openarchives.org

De acordo com [OAIa]: "Itens são objetos que constituem um repositório e que possuem metadados sobre um recurso que serão disseminados através dos registros em um formato específico". Na figura 3.6, temos uma ilustração dos três conceitos apresentados nessa definição. Aqui, a obra Monalisa é o recurso de informação. O item aqui corresponde ao conjunto de metadados que descrevem o recurso. Os metadados especificados são codificados em algum padrão (e.g. Dublin Core, MARC, Spectrum), e armazenados como registros de um banco de dados relacional (e.g. MySQL, Oracle), ou em uma forma menos usual, mas também eficiente (e.g. em arquivos textos ou em XML).

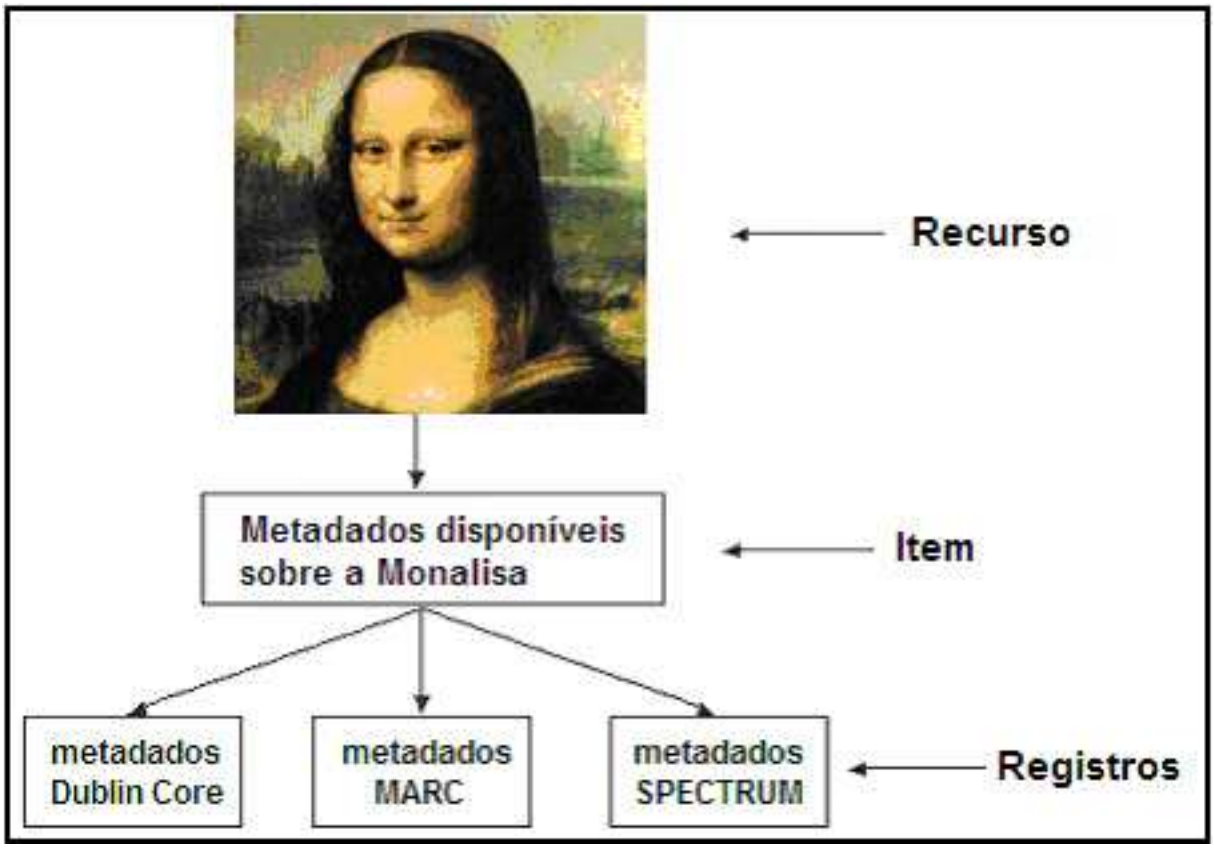

Figura 3.6 Recurso, item e registro.

Dentro de um dado repositório, cada item individual deve possuir um identificador único, usado para requisição dos metadados de um recurso. Este identificador único deve estar no formato do *oai-identifier* [OAId] e possui o seguinte padrão:

oai-identifier = oai ":"namespace-identifier ":"local-identifier

O termo **namespace-identifier** equivale ao domínio onde o repositório está localizado (e.g.: louvre.fr, cin.ufpe.br) e o **local-identifier** corresponde ao identificador que o recurso possui onde está armazenado fisicamente (e.g.: id, número seqüencial, etc.). Assim, o identificador único da obra de Leonardo da Vinci poderia ser:

oai:louvre.fr:789-Xy

Os registros dentro de um repositório devem ser formatados em três partes básicas [OAIa]:

- **Header:** É o cabeçalho de um item. Deve possuir os seguintes entes:
  identifier: um identificador único, já explicado anteriormente;
  datestamp: que informa a data de criação ou modificação do registro. Essa informação será útil para selecionar os registros para coleta.
  setSpec: zero ou mais conjuntos que compõem o repositório. Um item pode ou não pertencer a um ou mais conjuntos, a fim de que os coletores de dados consigam selecionar porções específicas dos repositórios.

- **Metadados:** Um formato específico de um item em algum padrão de metadados. Um repositório pode usar diversos padrões, mas, a fim de prover interoperabilidade, deve suportar no mínimo o formato Dublin Core (especificado como *oai_dc*).

- **about:** Esta parte é opcional de um registro e deve informar se ele está em conformidade com o esquema de XML adotado pelo repositório. Comunidades podem implementar esquemas de XML individuais que definem especificidades do conteúdo de seus repositórios. O about geralmente define rights statements (direitos de acesso que alguns itens possam possuir) e provenance statements (pode indicar de que repositório e quando o metadado foi coletado).

Na figura 3.7, mostramos um exemplo do recurso da Monalisa, codificado em XML, com todos os seus componentes.

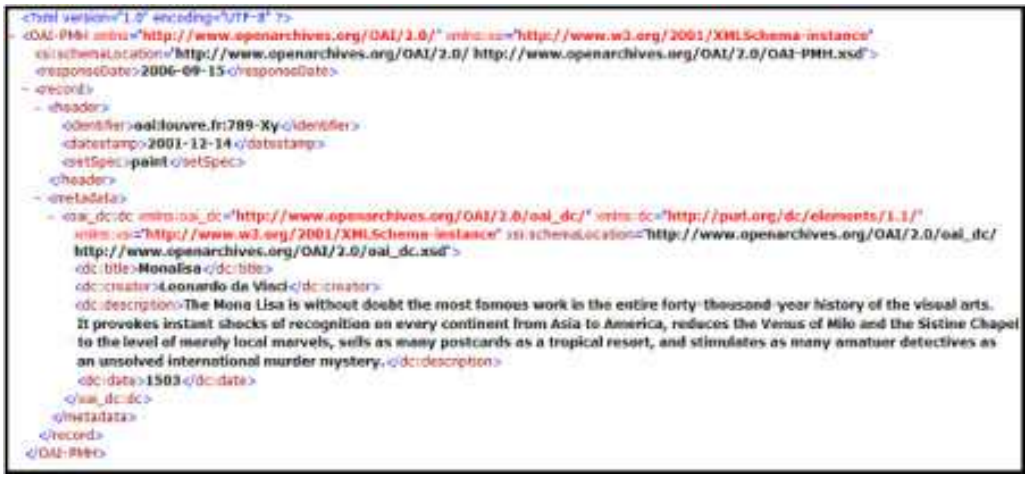

Figura 3.7 Codificação em XML de um item.

Por fim, para termos acesso ao Provedor de Dados via Internet, necessitamos de um servidor web (e.g..: Apache, IIS). Obviamente, também é necessária uma linguagem de programação (e.g.: PHP, Java) que implemente a exportação dos metadados do repositório em XML dentro do servidor escolhido e faça a conexão com a base de dados da aplicação.

3.2.2 Testes e Registro Oficial

Desde a primeira versão do protocolo OAI-PMH, em janeiro de 2001, o *Open Archives* vem mantendo um serviço de registros, tanto para Provedores de Dados, quanto para Provedores de Serviços. O registro para os Provedores de Dados não é obrigatório, afinal não há uma maneira para obrigar os desenvolvedores a fazerem isto [dSHL02].

Após completamente implementado, a melhor maneira de testarmos o Provedor de Dados é utilizando a ferramenta *Repository Explorer*, que será vista em detalhes na seção 3.5. O registro para a lista oficial de Provedores de Dados da OAI requer que o desenvolvedor forneça a URL básica do Provedor de Dados implementado (detalhes na seção 3.4). A partir daí, o OAI realiza um extensivo teste sobre o repositório e, havendo alguma irregularidade com o protocolo, é mandado ao requisitor do registro uma mensagem notificando o erro. Caso contrário, o Provedor de Dados implementado será adicionado na lista oficial dos repositórios OAI, sendo frequentemente checada sua conformidade [Forc].

3.3 Provedor de Serviços

Os Provedores de Serviços são sistemas que coletam e organizam os metadados dos repositórios OAI e os disponibilizam para o usuário final [Gar03]. Como já foi dito anteriormente, os Provedores de Serviços realizam requisições no formato HTTP para os Provedores de Dados, que por sua vez disponibilizam os dados em formato XML. Esses dados são coletados e normalmente armazenados dentro de uma base de dados (eg.: MySQL, Oracle). Essa coleta também é comumente chamada de *Havesting*. Assim, essa base pode ser disponibilizada para que usuários realizem consultas em um único acesso de documentos pertencentes a diversos repositórios.

Geralmente, essas coletas são realizadas de forma automática em um período fixo de tempo (e.g.: todas as madrugadas, semanalmente, etc.), para que as bases coletadas se mantenham atualizadas. Opcionalmente, as coletas podem ser feitas por um administrador do Provedor de Serviço.

Ao contrário do número de Provedores de Dados - que ultrapassa quinhentos - os Provedores de Serviços registrados pela OAI possuem um número bem mais reduzido, não chegando a trinta [2]. Na tabela 3.2 apresentamos uma lista dos mais conhecidos Provedores de Serviços OAI.

[2]Registered Service Providers - http://www.openarchives.org/service/listproviders.html

Tabela 3.2 Alguns Provedores de Serviços registrados

Nome do Provedor de Serviço	Descrição	URL
Arc	Serviço de Integração de metadados de acordo com o OAI que utiliza JDK 1.4, Tomcat 4.0 e Banco de Dados Relacional.	http://arc.cs.odu.edu/
BASE	Biblioteca Digital que integra repositórios OAI da área científica. Possui uma interface de busca aos documentos coletados, utilizando o software *FAST Search*.	http://www.base-search.net/
NCSTRL	Projeto que provê acesso a artigos científicos na área de Ciência da Computação.	http://www.ncstrl.org/
Sheet Music Consortium	Grupo de Bibliotecas que trabalham com o objetivo de criar uma coleção de músicas digitalizadas, integrando repositórios utilizando o OAI.	http://digital.library.ucla.edu/sheetmusic/

3.3.1 Pré-requisitos

Existem três pré-requisitos de infra-estrutura técnica que são necessários para a implementação de um Provedor de Serviços OAI [Forb]:

- **Um servidor conectado à Internet:** Por se tratar de um sistema Web, é pré-requisito básico que o Provedor de Serviço esteja conectado alocado em um servidor Web.
- **Um Sistema de Banco de Dados:** Esse sistema serve justamente para agregarmos os diversos metadados coletados do repositório em uma única base de dados. Vale ressaltar que estas bases de dados, geralmente são Banco de Dados relacionais, mas também podem ser arquivos XML ou arquivos texto.
- **Um Ambiente de Programação:** Este ambiente deve permitir que o Provedor de Serviços faça requisições HTTP para os repositórios, se comunique com a base de dados do repositório e inclua um parser XML, transformando o seu conteúdo em implementações SQL para inserção em sua base de dados.

3.3.2 Testes e Registro Oficial

Como já mencionamos, para testarmos um Provedor de Dados podemos utilizar a ferramenta Repository Explorer, que verifica se o seu repositório está de acordo com as especificações do OAI-PMH. No caso dos Provedores de Serviços, podemos testar a sua implementação (*havesting* e armazenamento) através dos repositórios registrados no site da OAI. Assim, testando um Provedor de Serviços com alguns desses repositórios trará a certeza que o sistema estará funcionando de acordo com o *Open Archives*.

Os Provedores de Serviços, assim como os Provedores de Dados, também podem ser oficialmente registrados no site da OAI. Uma vez testado, o responsável pelo sistema pode fazer uma

requisição à iniciativa, bastando informar algumas informações sobre o Provedor de Serviços (e.g.: nome do seu serviço, sua descrição, o público-alvo, a URL da página que é oferecido o serviço, uma lista dos Provedores de Dados que serão coletados, etc.) no momento do cadastro. O sistema será testado e, se estiver em conformidade com o protocolo, é adicionado à lista de Provedores de Serviços oficiais do OAI.

3.4 Protocolo OAI-PMH

Desde o primeiro lançamento do OAI-PMH, em Janeiro de 2001, o protocolo vem sendo considerado como a alternativa mais prática para a interoperabilidade entre Bibliotecas Digitais. O protocolo é baseado em uma simples estrutura, contendo duas partes principais: o Provedor de Dados e o Provedor de Serviços [dSHL02], ambos explicados anteriormente.

Os membros do OAI reconhecem que há limitações funcionais, se comparados com outros padrões de interoperabilidade, como o Z39.50 [LdS01]. De acordo com [Arm00], estratégias para interoperabilidade geralmente crescem em custos (dificuldade de implementação) com a adição de novas funcionalidades. O OAI-PMH não pretende substituir outras abordagens de integração, mas promover uma solução de simples implementação e disponibilização dos metadados para acesso integrado entre diferentes bases. Esta simplicidade de construção e acesso vêm sendo fatores determinantes para a grande ploriferação deste protocolo entre os mais diversos repositórios ao redor do mundo.

3.4.1 O *Harvesting*

Conforme já explicamos anteriormente, o protocolo OAI-PMH nos traz o conceito de *havesting*, que realiza a coleta dos metadados de diferentes repositórios. Um Provedor de Serviços, possuindo uma lista de repositórios registrados, realiza uma busca dentro desses Provedores de Dados, realizando solicitações para a coleta de seus metadados [Gar03].

Pode-se realizar esta coleta integralmente ou baseada em critérios. No caso de uma coleta integral, todos os metadados de um repositório são coletados e armazenados pelo Provedor de Serviços. Já o segundo tipo de coleta, possui os seguintes critérios para o *havesting*:

- **Baseado em data:** São coletados os metadados incluídos ou alterados após uma data especificada pelo Provedor de Serviço. Isso é feito verificando o elemento datestamp contido nos itens de metadados (ver seção 3.2.1).

- **Baseado em conjuntos:** São coletados os metadados de um conjunto específico do repositório coletado. Isso é feito verificando os valores do elemento setSpec (também na seção 3.2.1).

- **Híbrida:** Mescla os dois critérios acima. Assim, pode-se realizar a coleta dos metadados a partir de uma data qualquer dentro de um conjunto específico.

3.4.2 Os Verbos

Como dito, as requisições dos Provedores de Serviços são realizadas em formato HTTP, através de uma URL básica do repositório mantido pelo Provedor de Dados. Além da *URL básica* para coleta, é preciso identificar o que será coletado e como a mesma será realizada (i.e. identificar o tipo de *havesting*). Para isto, o OAI define seis verbos que especificam detalhes da coleta dos repositórios e alguns argumentos, a fim de refinar o *havester*.

Os verbos são especificados como parâmetros em uma requisição HTTP. Como exemplo, suponha um Provedor de Dados com a seguinte URL básica:

http://www.bibliox.org.br/pd/oai.php

Se quisermos, por exemplo, retornar todos os metadados do repositório, utilizamos o verbo *ListRecords*, e a requisição para o Provedor de Dados seria:

http://www.bibliox.org.br/pd/oai.php?verb=ListRecords&metadaPrefix=oai_dc

O outro argumento utilizado, o *metadataPrefix*, especifica em que padrão o metadado está formatado.

Como outro exemplo, considere a coleta dos metadados que foram inseridos no Provedor de Dados entre os dias 30 de janeiro de 2006 e 25 de outubro do mesmo ano. Para isso, são utilizados os argumentos*from* e *until*, do verbo *ListRecords*. A requisição ficaria da seguinte forma:

http://www.bibliox.org.br/pd/oai.php?verb=ListRecords&metadaPrefix=oai_dc
&from=2006-01-30&until=2006-10-25

Na tabela 3.3, encontramos os verbos do protocolo, suas descrições correspondentes e os argumentos que cada verbo deve e/ou pode utilizar [OAIb].

Tabela 3.3 Os verbos e seus argumentos

Verbo	**Descrição**	**Argumentos**
GetRecord	Recupera os metadados de um item individual de um repositório.	**identifier**. Obrigatório. Com ele, especificamos o identificador único (ver seção 3.2.1) do item de um repositório. **metadataPrefix**. Obrigatório. Especifica o padrão de metadados adotado que deve estar especificado no Provedor de Dados.
Identify	É usado para coletar informações sobre um repositório.	Não há argumentos.
ListRecords	Este verbo recupera os metadados de um repositório.	**from.** Opcional. Os dados coletados devem ser criados ou alterados a partir da data específica por este argumento. **until.** Opcional. Os dados coletados devem ser criados ou alterados até a data especificada pelo argumento. **metadataPrefix.** Já explicado anteriormente. **set.** Opcional. Especifica um conjunto, para o *havester* poder refinar a sua coleta. **resumptionToken.** Exclusivo. Argumento necessário quando os provedores utilizam o controle de fluxo na coleta dos metadados.
ListIdentifiers	É uma abreviação do ListRecords, que retorna apenas o *header* (ver seção 3.2.1) de um repositório.	**from.** **until.** **metadataPrefix.** **set.** **resumptionToken.**
ListMetadataFormats	Retorna os padrões de metadados utilizados em um repositório.	**identifier.** Opcional (apenas neste verbo). Retorna o padrão de metadados utilizado em um item específico.
ListSets	É utilizado para retornar a estrutura de um repositório, listando todos os conjuntos que compõe os metadados	**resumptionToken.**

Desta maneira, o protocolo OAI-PMH funciona com um Provedor de Serviços coletando informações dos diversos Provedores de Dados, através de verbos e argumentos, por requisições HTTP, como mostra a figura 3.8 [Forc].

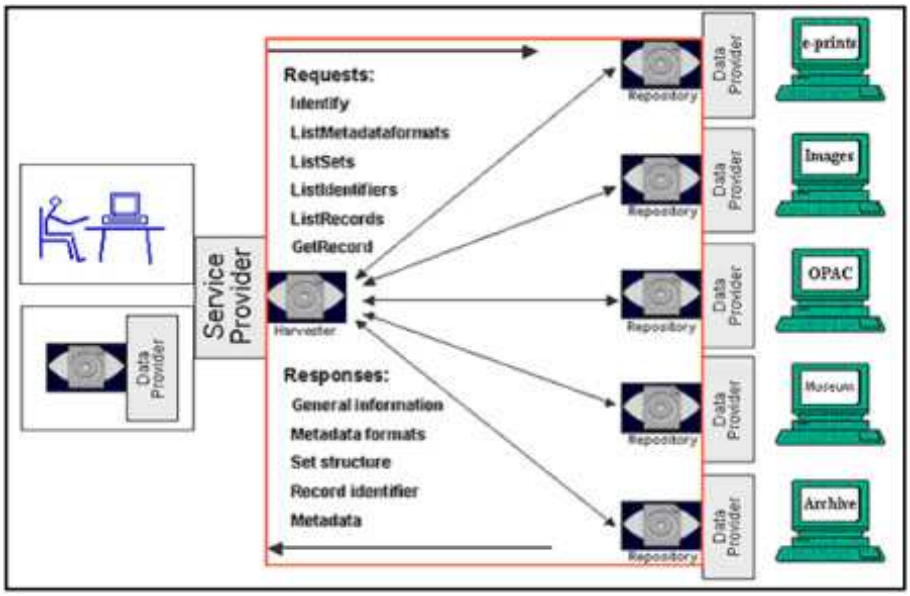

Figura 3.8 Funcionamento do Protocolo OAI-PMH através de seus verbos.

3.4.3 Requests e Responses

Como já foram mencionadas anteriormente, as respostas (*responses*) às requisições HTTP - com seus verbos e argumentos - são retornadas em formato XML. Abaixo, listaremos os seis verbos OAI, com exemplos de um request HTTP e o seu response XML de alguns Provedores de Dados cadastrados no site da OAI.

GetRecord

Se quiséssemos retornar um item específico do repositório *Analytical Sciences Digital Library*[3], poderíamos solicitar a seguinte requisição:

http://www.asdlib.org/oai/oai.php?verb=GetRecord
&identifier=oai:asdlib.org:asdl002941&metadataPrefix=oai_dc

A Figura 3.9 mostra o *response* correspondente:

[3]Analytical Sciences Digital Library - http://www.asdlib.org/

Figura 3.9 Exemplo do verbo GetRecord.

Identify

A requisição seguinte nos retorna a identificação do Provedor de Dados *Auburn University Digital Library*[4], cujo response pode ser verificado na figura 3.10:

http://content.lib.auburn.edu/cgi-bin/oai.exe?verb=Identify

Figura 3.10 Exemplo do verbo Identify.

[4]Auburn University Digital Library - http://content.lib.auburn.edu/

ListRecords

Para requisitarmos os registros da Biblioteca *A Celebration of Women Writers*[5], basta usarmos este verbo. A figura 3.11 corresponde ao response do seguinte request (com os argumentos opcionais *from* e *until*):

http://digital.library.upenn.edu/webbin/OAI-celebration?verb=ListRecords &metadataPrefix=oai_dc&from=2006-05-06&until=2006-05-20

Figura 3.11 Exemplo de um ListRecords.

[5] A Celebration of Women Writers - http://digital.library.upenn.edu

ListIdentifiers

Para descobrirmos os identificadores do Provedor de Dados *Elektronisches Publikationsportal der Österreichischen Akademie*[6], utilizamos a seguinte requisição:

http://hw.oeaw.ac.at/oai?verb=ListIdentifiers&metadataPrefix=oai_dc
&from=2006-05-20&until=2006-06-120

E o seu respectivo *response* é mostrado na figura 3.12:

Figura 3.12 Exemplo do verbo ListIdentifiers.

ListMetadataFormats

Com este verbo, poderíamos saber que padrões de metadados são utilizados nos repositórios, como na *Indiana University Digital Library Program*[7], através da requisição:

http://oai.dlib.indiana.edu/phpoai/oai2.php?verb=ListMetadataFormats

Nesta Biblioteca Digital encontramos dois formatos, conforme mostra a figura 3.13.

Figura 3.13 Exemplo de ListMetadataFormats.

[6]Elektronisches Publikationsportal der Österreichischen Akademie - http://hw.oeaw.ac.at/
[7]Indiana University Digital Library Program - http://oai.dlib.indiana.edu/

ListSets

Por fim, podemos saber quais os conjuntos que são utilizados no repositório *The University of Texas at Austin Libraries*[8] através da requisição:

http://www.lib.utexas.edu/oai/oai2.php?verb=ListSets

Os conjuntos que organizam os metadados desta Biblioteca Digital podem ser verificados na figura 3.14.

Figura 3.14 Exemplo de ListSets.

3.4.4 Erros e Condições de Exceção

Na subseção anterior, especificamos como procedemos a requisição dos Provedores de Dados no protocolo através dos verbos da OAI. Entretanto, se alguma dessas requisições não estiver bem formatada, o repositório em questão deve levantar uma exceção.

Na tabela 3.4 encontramos os possíveis erros adotados pelo OAI-PMH e em que verbo os mesmos podem ser encontrados [OAIa]:

[8]The University of Texas at Austin Libraries - http://www.lib.utexas.edu/

Tabela 3.4 Relação dos erros do OAI-PMH

Erro	**É chamado quando**	**Verbos aplicáveis**
badArgument	Algum argumento ilegal é utilizado Algum argumento obrigatório não é utilizado Argumentos repetidos Valores de argumentos com sintaxe ilegal	Todos os verbos
badResumptionToken	O valor do *resumptionToken* é inválido	ListIdentifiers ListRecords ListSets
badVerb	O valor do verbo é inválido Não existe o argumento *verb*	
cannotDisseminateFormat	O formato do metadados indicado pelo argumento *metadataPrefix* não é suportado pelo repositório	GetRecord ListIdentifiers ListRecords
idDoesNotExist	O valor do argumento *identifier* é desconhecido do repositório	GetRecord ListMetadataFormats
noRecordsMatch	A combinação dos valores dos argumentos *from*, *until*, *set* e *metadataPrefix* retornam uma lista vazia	ListIdentifiers ListRecords
noMetadataFormats	Quando não há formatos de metadados disponíveis para um item específico	ListMetadataFormats
noSetHierarchy	O repositório não suporta conjuntos	ListSets ListIdentifiers ListRecords

Já mencionamos anteriormente, mas vale salientar que a resposta a possíveis erros é também retornada em formato XML. Por exemplo, se especificarmos um verbo inexistente a um Provedor de Dados através da requisição:

http://oai.dlib.indiana.edu/phpoai/oai2.php?verb=ListarRegistros

Com esta requisição a um verbo inexistente (ListarRegistros), o repositório deve levantar uma exceção através do erro *badVerb*, conforme mostrado na figura 3.15:

Figura 3.15 Exceção a uma requisição através do erro badVerb.

3.4.5 Controle de Fluxo

Em alguns casos, quando um Provedor de Serviço requisita os metadados de um repositório, a lista com todos os registros pode ser muito grande e carregá-la de uma só vez pode causar

quedas se o servidor não for muito robusto. Desta maneira, particionar o resultado de uma coleta de metadados pode significar um ganho em processamento e diminuição do risco de um servidor falhar no momento da coleta desses dados [dSHL02].

Este particionamento é realizado da seguinte maneira:

- Um repositório responde a uma requisição de um Provedor de Serviços através de uma lista incompleta dos metadados e um *resuptionToken* (uma variável que indica os restantes dos itens que ainda restam no repositório).

- Para coletar uma lista completa, o coletor de dados necessita de várias requisições com *resuptionToken* como argumento. A lista completa consiste na concatenação de listas incompletas de uma seqüência de requisições.

Na figura 3.16 representamos graficamente como se dá o funcionamento desta importante funcionalidade [Forb].

Figura 3.16 Controle de Fluxo do OAI.

3.5 Ferramentas e Bibliotecas Digitais OAI

Muitas instituições ligadas à iniciativa do *Open Archives* desenvolveram ferramentas com o auxílio de cooperar com o advento do OAI, facilitando outras instituições a implementar Bibliotecas Digitais em conformidade com o protocolo. No site da OAI, encontramos um total de trinta ferramentas registradas como oficiais [9]. Daremos uma breve explicação das que julgamos as mais importantes além de citarmos algumas Bibliotecas Digitais que trabalham em conformidade com o *Open Archives*.

[9]PMH Tools - http://www.openarchives.org/pmh/tools/tools.php

3.5.1 Repository Explorer

O Repository Explorer é uma ferramenta Web que testa o pleno funcionamento de um Provedor de Dados, verificando se o Provedor de Dados suporta todos os verbos e argumentos do OAI-PMH. Criada pela *Vermont University*[10], a ferramenta suporta testes manuais e automáticos. No modo de testes automático, uma série de requisições do protocolo são realizadas, com parâmetros legais ou não, contra o arquivo a ser testado. Já no modo manual, o software permite que o usuário defina quais requisições e parâmetros serão testados, através de formulários HTML [Sul01].

O Repository Explorer suporta diversos idiomas como: Chinês, Inglês, Espanhol, Francês Alemão, Coreano e Português. A interface principal do sistema pode ser vista na figura 3.17.

Figura 3.17 A ferramenta de testes Repository Explorer.

3.5.2 OAI-Cat

Desenvolvido pela *Online Computer Library Center*[11] , o OAI-Cat[OCL05] se trata de uma ferramenta *Open Source* baseada em aplicação *Web Java Servlet*. Ela provê um *framework* para Provedores de Dados em conformidade com o protocolo OAI-PMH. A ferramenta pode ser customizada para trabalhar com diversos repositórios, contudo, algumas interfaces em Java devem ser implementadas para tal.

[10]The University of Vermont - http://www.uvm.edu/
[11]Online Computer Library Center - http://www.oclc.org/

3.5.3 Arc

A ferramenta Arc [LMZN01] é um Provedor de Serviços para Bibliotecas Digitais. Desenvolvido pela *Old Dominion University*[12] , ela realiza o havesting entre diversos repositórios e os armazena em uma base de dados relacional, dando a opção de utilização do Oracle e do MySQL.

A sua arquitetura é baseada em Java Servlets, é independente e pode trabalhar com qualquer *Web Server*. Na figura 3.18[LMZN01] encontramos a arquitetura deste Provedor de Serviços, destacando outros componentes como a interface com o usuário e o *havesting*.

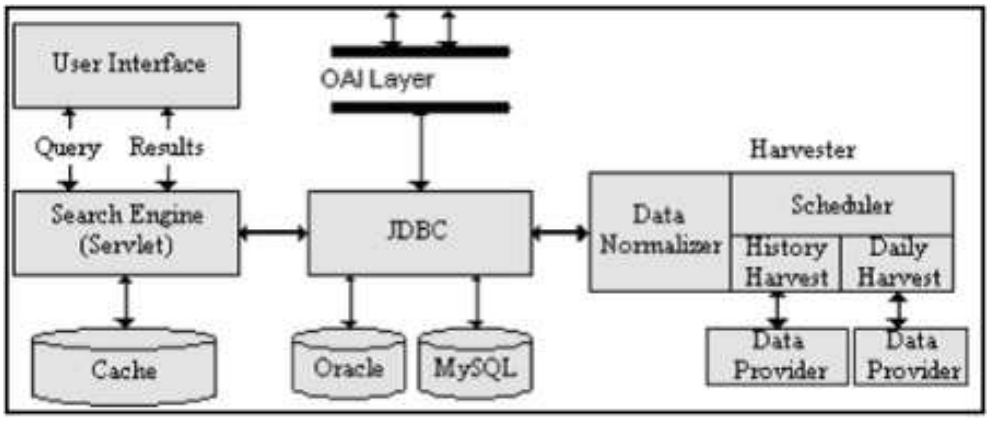

Figura 3.18 A Arquitetura do Arc.

A ferramenta pode realizar a coleta dos dados de quatro formas:

- **Coleta de todos os metadados históricos:** Realiza a coleta dos metadados inseridos no repositório que não tenham sido coletados e que não sofreram alteração a partir de uma determinada data informada.
- **Coleta de todos os metadados novos:** Realiza a coleta de metadados que já tenham sido coletados, mas que sofreram modificações nos seus Provedores de Dados.
- **Coleta individual de um metadado histórico:** Realiza a coleta de apenas um recurso dentro de repositório.
- **Coleta individual de um novo metadado:** Realiza a coleta de um recurso no repositório que já tenha sido coletado, mas que sofrera atualização no repositório de origem.

De certa forma, a quantidade de maneiras para realizar as coletas pode parecer atrativo, à primeira vista. Entretanto, deve-se oferecer ao usuário uma maneira mais usual de realizar essas coletas, fornecendo mecanismos de auxílio na integração dos metadados. Além do mais, a ferramenta em questão não implementa o controle de fluxo, requisito de grande funcionalidade para sistemas OAI.

[12]Old Dominion University - http://www.odu.edu/

O Arc ainda oferece aos seus usuários um serviço de consulta dos dados contidos na sua base. Essa busca pode ser simples ou avançada.

3.5.4 E-Prints

O E-Prints é uma das ferramentas mais utilizadas para implementações de repositórios digitais. É compatível com o protocolo OAI-PMH, já disponibiliza versões bastante estáveis e possui uma boa documentação sobre o software [Brac].

Um dos pontos fortes da ferramenta é a sua característica de permitir o *self-archiving*, ou seja, depósitos de documentos eletrônicos ainda em construção pelos próprios autores em repositórios digitais. Implementado em linguagem Perl, os usuários podem alimentar o repositório com arquivos que são armazenados em uma base de dados MySQL. Além de funcionar como um Provedor de Dados, a ferramenta ainda oferece um *harvester*, que coleta informações de outros repositórios que utilizam o E-Prints.

Por ser desenvolvido com a visão voltada para abrigar artigos científicos, o seu uso para outros tipos de documentos pode tornar-se um trabalho de custos não compensáveis. Outra desvantagem do E-Prints deve-se ao fato de sua interface não ser muito amigável: o depósito de um artigo conta com muitos formulários diferentes e seu mecanismo de busca oferece muitas opções ao usuário, dificultando o trabalho do usuário [Brac]. Além do mais, a ferramenta só coleta metadados de repositórios que estejam oficialmente registrados no *Open Archives*.

3.5.5 Biblioteca Digital *PubMed Central*

A figura 3.19 apresenta o site da Biblioteca Digital PubMed Central[Cen]. Criada em fevereiro do ano 2000 pelo *National Center for Biotechnology Information*[13] , Estados Unidos, abriga documentos científicos sobre biomédica e ciência da vida. O sistema em questão está de acordo com a iniciativa do *Open Archives*, entretanto, ele apenas é um Provedor de Dados. Além do mais, os recursos eletrônicos não estão disponíveis para o *havester*, só estando restando os seus metadados para coleta. Outra característica da Biblioteca Digital é possuir apenas documentos do tipo texto.

[13]National Center for Biotechnology Information - http://www.ncbi.nlm.nih.gov/

Figura 3.19 Biblioteca Digital PubMed Central.

3.5.6 *Bielefeld Academic Search Engine*

Um outro projeto bastante interessante que trabalha de acordo com o OAI trata-se da Biblioteca Digital denominada *Bielefeld Academic Search Engine* (BASE) [PS05], Alemanha. Desenvolvida pela *Bielefeld University Library*[14] a BASE é um Provedor de Serviços que coleta documentos científicos das mais diversas áreas de conhecimento, em 283 repositórios, abrigando mais de quatro milhões de registros.

Além de abrigar uma quantidade significativa de metadados, o sistema possui ainda um sistema de busca simples e eficiente, mostrado na figura 3.20. No canto direito da tela, podemos observar que o usuário tem a opção de refinar a sua consulta ou ordenar o resultado da mesma segundo alguns critérios (e.g.: nome do autor, tamanho do documento, data de publicação, etc.).

[14]Bielefeld University Library - http://www.ub.uni-bielefeld.de

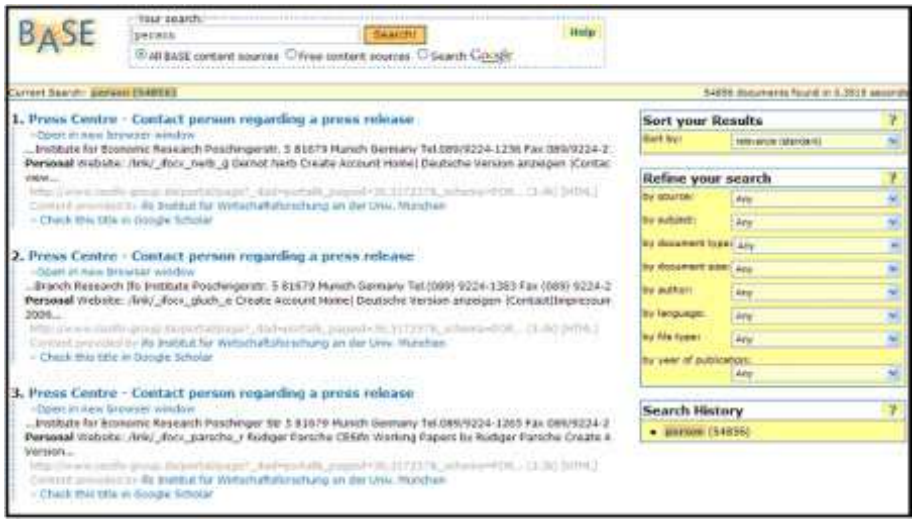

Figura 3.20 Biblioteca Digital BASE.

3.5.7 Perseus Lookup

A terceira Biblioteca Digital apresentada é a *Perseus Lookup*, projeto que reúne fontes sobre o estudo da humanidade. Este sistema também se trata de um Provedor de Serviços OAI, reunindo metadados de 27 repositórios diferentes.

Uma funcionalidade interessante desta Biblioteca Digital é a possibilidade de realizarmos a busca de informações pertencentes a um Provedor de Dados específico. Na figura 3.21, que mostra a página de buscas do sistema, o campo *Any collection* contém a lista dos repositórios existentes.

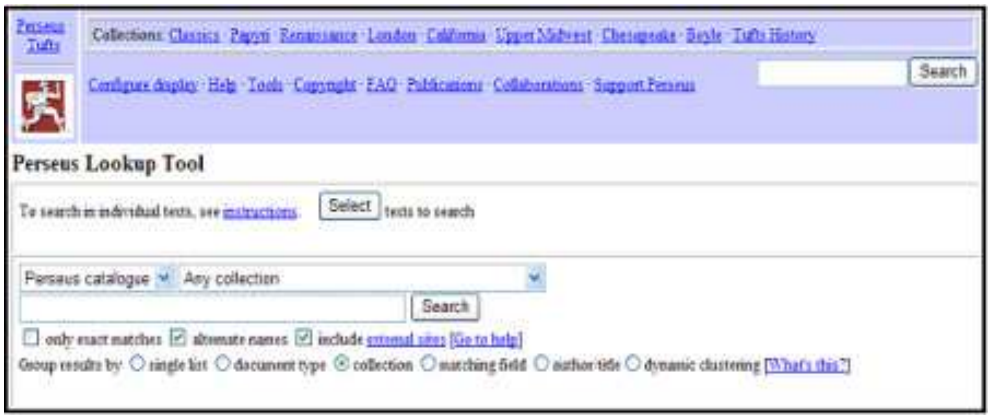

Figura 3.21 Biblioteca Digital Perseus.

O Perseus, entretanto, apresenta duas desvantagens sérias: fraca usabilidade e sistemas de buscas com performance baixa. Primeiramente, o retorno da consulta não informa os metada-

dos de cada documento. Ora apresenta um texto informando cada documento, ora *links* externos, mas que não vão direto ao documento informado. Quanto à baixa performance, realizamos uma série de consultas na Biblioteca Digital e cada um delas não durou menos do que dois minutos, um tempo inaceitável para sistemas Web atualmente.

3.5.8 OAIster

Finalmente, a última Biblioteca Digital filiada ao Open Archives apresentada é o OAIster [OAIe], uma das melhores do gênero. Desenvolvida pela *University of Michigan Digital Library Production Service*[15] tem o objetivo de reunir trabalhos acadêmicos e facilitar o acesso a esses recursos digitais. O OAIster recolhe metadados de 704 instituições, acumulando mais de 9 milhões de documentos para consulta dos usuários.

Na interface de busca do sistema, o usuário pode definir como o resultado será ordenado (e.g.: por título, autor, etc.), e escolher que tipo de mídia necessita (e.g.: áudio, vídeo, texto, imagem), caracterizando-se uma Biblioteca Digital multimídia. Na figura 3.22 demonstramos o resultado de uma consulta no OAIster. Como se pode perceber, os metadados dos documentos são apresentados separadamente, de maneira que o usuário pode distinguir claramente cada um deles. Além do mais, um recurso interessante aparece no canto esquerdo da tela: os resultados também são separados por instituições.

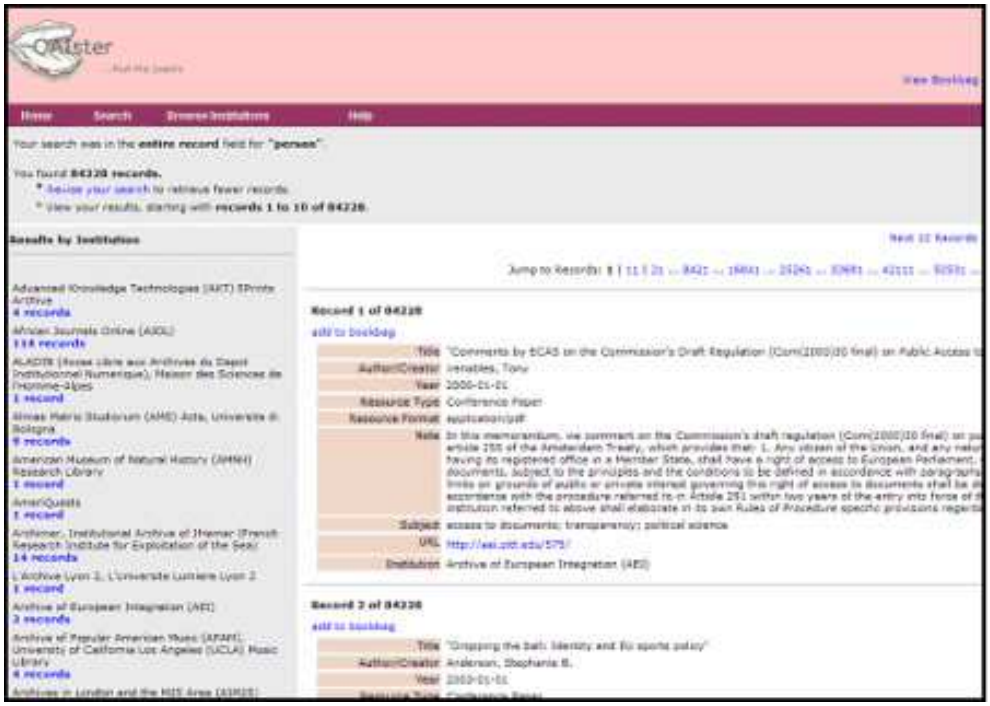

Figura 3.22 Biblioteca Digital OAIster.

Entretanto, nem sempre o documento é retornado junto com os seus metadados. Algumas

[15]University of Michigan Digital Library Production Service - http://www.umdl.umich.edu/

vezes, a URL (metadado que remeteria ao documento original) remete o usuário a uma página Web onde o recurso não se encontra. Assim como nos outros Provedores de Serviços mostrados nesta subseção, o OAIster também não possui a função de um Provedor de Dados, perdendo uma excelente oportunidade de disponibilizar milhares de documentos já coletados.

3.6 Critérios de Avaliação

A partir dos sistemas analisados, podemos definir alguns critérios de avaliação que consideramos relevantes que Bibliotecas Digitais que trabalhem com interoperabilidade de dados, especificamente com o OAI, possuam. São eles:

- **Ser um Provedor de Dados:** A fim de permitir uma maior disseminação do conteúdo ao alcance de todos, toda Biblioteca Digital deveria disponibilizar os seus dados para coleta de outros sistemas (evidentemente, respeitando os direitos autorais de cada documento). Ou seja, toda Biblioteca Digital deveria ser um Provedor de Dados. Inclusive, os Provedores de Serviços que, como vimos, chegam a armazenar milhões de metadados oriundos de vários repositórios, deveriam também ser um Provedor de Dados.

- **Coletar não só os metadados, mas o recurso também:** Da mesma forma que definimos a visualização do conteúdo completo de um recurso como um dos critérios para avaliação de Bibliotecas Digitais, seria interessante que os sistemas em questão também trabalhassem com o conteúdo completo. Coletar milhões de itens sem disponibilizar a informação em si, muitas vezes não interessa ao usuário.

- **Realizar a busca em um repositório específico:** A Biblioteca Digital que trabalha como um Provedor de Serviços coleta dados de diversos repositórios. Talvez interesse ao usuário de uma Biblioteca Digital consultar dados apenas de um desses repositórios. Por isso, essa funcionalidade deveria ser implementada no sistema.

No capítulo anterior, definimos quatro critérios que Bibliotecas Digitais deveriam possuir para atenderem o usuário da melhor maneira possível. Neste capítulo, especificamos ainda mais o critério da integração de dados e, a partir dele, definimos mais três critérios para sistemas OAI. Agora, relacionamos as ferramentas e Bibliotecas Digitais mostradas anteriormente com todos os sete critérios definidos para este tipo de sistema, apresentados na tabela 3.5.

A partir da tabela apresentado, percebemos que os Provedores de Serviços parecem preocupar-se apenas em coletar os metadados dos mais diversos repositórios. Contudo, elas também deveriam realizar o seu papel de divulgador da informação. Segundo [Braa]: "Uma Biblioteca Digital deve, dentro do possível, assemelhar-se mais com uma biblioteca tradicional, para que o usuário sinta-se mais familiar com o ambiente". Entendemos, dessa maneira, que além de possuir um coletor de dados, este tipo de sistema deve disponibilizar as informações completas, afim de que o usuário possa ter acesso a todo o recurso eletrônico de forma simples e eficiente.

Tabela 3.5 Critérios de avaliação de ferramentas e Bibliotecas Digitais OAI

	OAI-Cat	**E-Prints**	**Arc**	**PubMed Central**	**BASE**	**Perseus**	**OAIster**
Sistema de Busca	Não	Sim	Sim	Sim	Sim	Sim	Sim
Visualização dos Documentos	Não	Não	Não	Não	Não*	Não	Não*
Biblioteca Multimídia	Não	Não	Não	Não	Não	Não	Sim
Integrador de Dados	Não	Sim	Sim	Não	Sim	Sim	Sim
Provedor de Dados	Sim	Sim	Sim	Sim	Não	Não	Não
Coleta o recurso eletrônico	Não	Não	Não	Não	Não	Não	Não
Consulta em um repositório específico	Não	Sim	Sim	Sim	Sim	Sim	Sim

*disponibiliza um link que nem sempre é o documento

3.7 Considerações Finais

A união de informação devidamente estruturada com a facilidade de sua coleta e disponibilização tornam o *Open Archives Initiative* cada vez mais poderoso para a interoperabilidade entre Bibliotecas Digitais. O crescimento de sua utilização em diversos projetos ao redor do mundo deve-se principalmente ao fato da facilidade e eficiência do seu protocolo, o OAI-PMH. Hoje já existem diversas Bibliotecas Digitais que possuem vários documentos em suas bases, como o OAIster e o BASE, que juntos, somam quase 15 milhões de informações coletadas de diferentes repositórios.

Entretanto, por se tratar de uma tecnologia ainda recente (a primeira versão do protocolo foi lançada em 2001), muitas são as barreiras a serem enfrentadas. A principal delas é tratar um Provedor de Serviços apenas como um agente que coleta informações de repositórios OAI. Como o próprio nome sugere, além desta funcionalidade, ele deve prover ao usuário um serviço de Biblioteca Digital: prover acesso à informação desejada de forma clara, objetiva, simplificada e completa.

Desta maneira, o objetivo dos capítulos seguintes é a apresentação de uma tentativa de solucionar algumas lacunas deixadas por estes tipos de sistemas, principalmente no que diz respeito à interoperabilidade de dados.

CAPÍTULO 4

O Clio-i

Na maior parte das vezes, lembrar não é reviver, mas refazer, reconstruir, repensar, com imagens e idéias de hoje, as experiências do passado. A memória não é sonho, é trabalho.

—ECLÉA BOSI

Nos dois capítulos anteriores, apresentamos alguns critérios de avaliação que julgamos essenciais para um sistema de Bibliotecas Digitais. Critérios como visualização do documento, suporte a informação de várias mídias e sistemas de busca qualificados dificilmente são encontrados nestes tipos de sistema. Dentre esses critérios, destacamos a interoperabilidade de dados entre suas bases, disponibilizando ao usuário informações coletadas de diversos repositórios em um único ponto de acesso.

O objetivo deste capítulo é apresentar o Clio-i, um sistema para Bibliotecas Digitais que agrega todos os critérios de avaliação relatados nos capítulos passados. Para isso, citaremos algumas motivações para a construção deste tipo de sistema, mostraremos a arquitetura, seus módulos e funcionalidades. Quanto à interoperabilidade entre as bases, julgamos necessários realizar algumas modificações no protocolo OAI-PMH para o seu melhor aproveitamento, que serão descritos também nas próximas seções.

4.1 Histórico

Todo estudo a respeito de Bibliotecas Digitais para o presente trabalho foi iniciado no Liber , laboratório que desenvolve investigações no campo do gerenciamento eletrônico do conhecimento da Universidade Federal de Pernambuco. Para este fim, constrói, em ambientes controlados, repositórios e ferramentas que permitem a disponibilização, gerenciamento e pesquisa de conteúdos de formato digital [Gal05].

Em meados do ano de 2004, deu início no Laboratório Liber[1] um estudo sobre Bibliotecas Digitais e os serviços que as mesmas deveriam oferecer. Os primeiros testes foram realizados com o projeto Ultramar[2] que abrigava milhares de documentos da época do Brasil Colônia. O projeto consistia em uma Biblioteca Digital construída a partir da digitalização do *Fundo Arquivístico José Antônio Gonsalves de Mello*. Esse arquivo é constituído por um conjunto de 60.000 documentos microfilmados sobre a história do Brasil Colônia, particularmente, sobre Pernambuco e outros estados do Nordeste. A maior parte desses documentos, datados do século

[1]Laboratório Liber - http://www.liber.ufpe.br

[2]Biblioteca Digital Ultramar - http://www.liber.ufpe.br/ultramar

XVI e XVII, foi coletada pelo pesquisador José Antônio Gonsalves de Mello ainda na década de 1950, em suas pesquisas no Arquivo Histórico Ultramarino de Lisboa.

Esta primeira versão do sistema tratava-se apenas de uma consulta a uma base com os metadados sobre o acervo da época, disponibilizando a primeira página do documento.

Mesmo com poucos recursos e funcionalidades, o sistema já despertou interesse da comunidade científica e de diversas instituições nacionais e internacionais, justamente pela escassez de Bibliotecas Digitais que abrigassem documentos pesquisáveis.

Com a grande demanda percebida e a visão voltada para correta disseminação da informação e a preservação do documento digital, foi iniciado o estudo sobre um sistema que pudesse ampliar os horizontes na área. Nessa época, tal sistema abrigava essencialmente documentos históricos. Com o apoio da Fundação Joaquim Nabuco[3] e da Fundação Gilberto Freyre[4], foram identificados alguns requisitos básicos para Bibliotecas Digitais para Documentos Históricos, como sistema de busca, visualização do documento histórico por completo e manipulação do mesmo para uma melhor visualização. Neste mesmo levantamento, foi esquematizado a estruturado da base de dados que abrigariam as informações dos acervos e quais metadados poderiam identificar os documentos históricos. Nossa intenção era criar uma Biblioteca Digital que atendesse a qualquer documentação histórica, e desta maneira, abrigar diversos projetos para este fim.

Neste processo de criação da segunda versão do sistema, foi criada a Biblioteca Digital Gilberto Freyre. Essa biblioteca foi construída inicialmente com a série de correspondências trocadas entre o escritor Gilberto Freyre e diversas personalidades da área política, cultural e social do Brasil e do exterior. Dentre elas, citamos Jorge Amado, Carlos Drummond de Andrade, Assis Chateaubriand, Getúlio Vargas, Câmara Cascudo e Darcy Ribeiro.

Esta segunda versão abrigou ainda o projeto Ultramar, anteriormente citado. Desta maneira, toda a base antiga do projeto foi migrada para esta nova versão do sistema, que pode ser verificada na figura 4.1.

Figura 4.1 O Projeto Ultramar atualmente.

[3]Fundação Joaquim Nabuco - http://www.fundaj.gov.br
[4]Fundação Gilberto Freyre - http://www.fgf.org.br

Os estudos no âmbito de Bibliotecas Digitais avançaram e logo percebemos que o sistema em que trabalhávamos poderia ser revisto e aplicado a qualquer tipo de fundo documental, não apenas a documentos históricos. Ou seja, nossa intenção agora era criar uma Biblioteca Digital de tema geral que estivesse de acordo com as funcionalidades que um sistema desse porte necessita.

A partir desse momento, é que atingimos o estado atual do presente trabalho. Após um estudo detalhado dos serviços que algumas Bibliotecas Digitais oferecem a seus usuários - discutido amplamente no capítulo 2 -, encontramos algumas carências que tentaremos solucioná-las.

Além disso, um assunto amplamente discutido na área de Bibliotecas Digitais é a interoperabilidade entre diversas bases através da Internet. Pesquisas nesse âmbito foram realizadas e as detalhamos também no capítulo 2. Dentre as iniciativas mais utilizadas para este fim, encontramos o *Open Archives Initiative*, discutido com detalhes no capítulo 3 e que servirá como base para o desenvolvimento da nova versão da Biblioteca Digital.

Após os estudos realizados, propomos o Clio-i, um sistema que oferece aos usuários serviços qualificados que uma Biblioteca Digital deve possuir e, principalmente, realiza a coleta de metadados de diversas bases diferentes que estejam de acordo com o protocolo OAI-PMH.

4.2 Funcionalidades Principais

Conforme já mencionado no capítulo 2, uma Biblioteca Digital é uma das formas mais avançadas e complexas de sistema de informação, pois envolve, dentre outras características, preservação do documento digital, disseminação eficiente do conteúdo, serviço de informação multimídia e gerenciamento de bases distribuídas. Entretanto, também foi apresentado alguns sistemas na área e nenhum deles atende todos os requisitos necessários.

Pretendemos assim, com o Clio-i, preencher essa lacuna na área de Bibliotecas Digitais, atendendo a algumas funcionalidades, na qual destacamos as principais:

- **Sistema de Recuperação da Informação:** O módulo de recuperação de informação do Clio-i é responsável por tratar as consultas dos usuários da biblioteca. Cada documento é indexado pelos metadados que serão detalhados mais à frente. A recuperação de documentos é realizada a partir de buscas baseadas em palavras-chave nesses campos, e a ordem de apresentação dos documentos recuperados em uma dada consulta é determinada pela sua relevância.

- **Suporte a Documentos Multimídia:** Com o advento tecnológico, a informação digital nos dias de hoje encontra-se em diferentes formatos. Por isso, o Clio-i deve suportar arquivos do tipo texto, áudio, vídeo e imagem.

- **Visualização dos Documentos:** Com raras exceções, uma Biblioteca Digital nos dias de hoje disponibiliza todo o conteúdo do documento. O Clio-i deve prover ao usuário essa funcionalidade, disponibilizando, sempre que possível, a informação por completo.

- **Manipulação dos Documentos:** O Clio-i também deve permitir que o usuário possa realizar algumas manipulações em um documento específico, como aumentar ou diminuir,

clarear ou escurecer, girar verticalmente ou horizontalmente, entre outros que serão citados mais adiante.

Essas funcionalidades mencionadas são os serviços oferecidos por uma Biblioteca Digital ao seu usuário. Entretanto, o grande diferencial do Clio-i é o seu poder de interoperabilidade com outros repositórios na Internet.

Para tal atividade, o sistema utiliza a iniciativa do *Open Archives*, já explicada no capítulo anterior. Desta maneira, qualquer repositório que atenda as exigências do protocolo OAI-PMH pode ter seus dados coletados pelo Clio-i. Por conseguinte, o Clio-i irá coletar os dados de vários repositórios, armazená-los em uma base de dados única e disponibilizá-los através dos serviços que serão oferecidos para o usuário, citados anteriormente.

Assim sendo, dentro do contexto da integração de Bibliotecas Digitais - especificamente utilizando o *Open Archives* - podemos citar as características principais que o Clio-i propõe atender:

- **O Integrador também deve expor seus dados:** Os sistemas avaliados que utilizam o protocolo OAI-PMH para a coleta dos dados não os expõem. Acreditamos que a exposição de dados já coletados agregaria um grande valor à disseminação eficiente da informação através da Internet. Desta maneira, o Clio-i também deverá expor dados armazenados que foram coletados de diversas outras fontes de dados.

- **Extensão do protocolo OAI-PMH:** Acreditamos que o protocolo da *Open Archives*, o OAI-PMH, atende a uma grande demanda na área da interoperabilidade entre repositórios digitais, por razões citadas no capítulo anterior. Entretanto, para que algumas necessidades que entendemos serem de alta relevância para um sistema desse porte fossem atendidas, precisávamos estender o protocolo. A justificativa e as modificações necessárias serão vistas mais adiante.

Na figura 4.2 ilustramos, em alto nível, o fluxo de informações do Clio-i. Nela, é apresentando a relação com os usuários, administradores e sua base de dados. Além do mecanismo de exportação e coleta das informações de outros repositórios OAI.

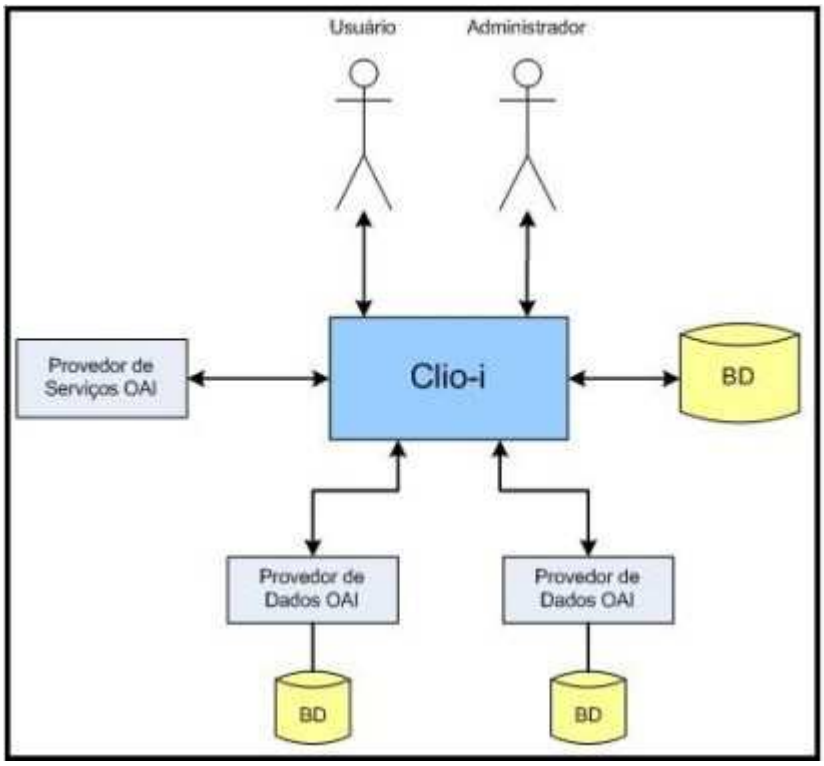

Figura 4.2 Fluxo de informações no Clio-i.

Primeiramente, de acordo com a figura 4.2, percebemos que o Clio-i se comunica com uma base de dados qualquer, que tem a função de armazenar todas as informações coletadas dos diversos Provedores de Dados OAI. Essas informações coletadas são apresentadas aos usuários através de alguns serviços (e.g. Recuperação de Informação, Visualização do Documento, etc.), todos citados anteriormente. Outro ator que interage com o sistema é o Administrador, que tem a função de gerenciar os repositórios que serão futuramente coletados pelo Clio-i. Por fim, o Clio-i também exporta os dados depositados na sua base, que podem ser coletados por qualquer Provedor de Serviço OAI.

Após uma descrição inicial do projeto Clio-i, analisaremos o mesmo com mais detalhes a partir das próximas seções.

4.3 Arquitetura do Sistema

Após os estudos realizados e apresentados nos dois capítulos anteriores, definimos uma arquitetura que acreditamos atender os requisitos de uma Biblioteca Digital com capacidade de integrar diferentes repositórios e fornecer serviços de qualidade ao usuário final.

Primeiramente, ressaltamos que o Sistema Clio-i conta com quatro módulos principais, todos apresentados na arquitetura geral do sistema, figura 4.3.

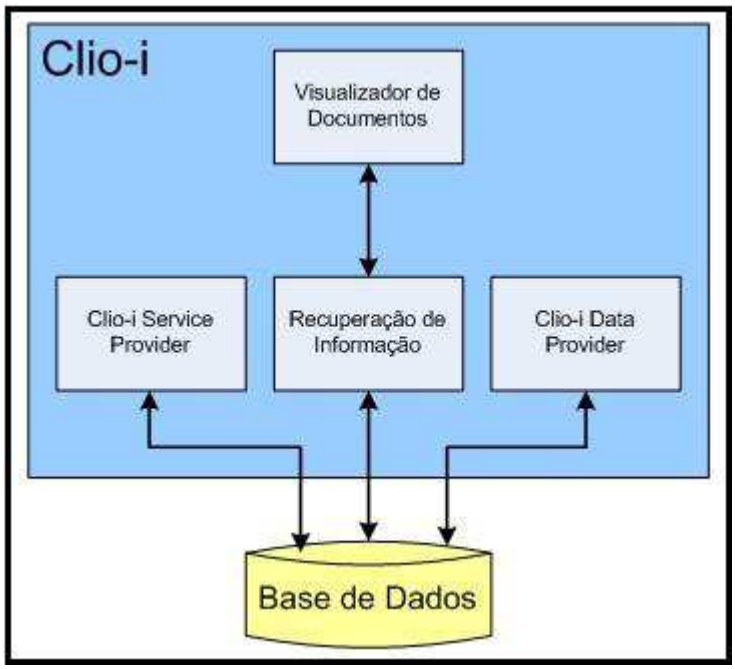

Figura 4.3 Arquitetura Geral do Sistema.

Todos os módulos são explicados a seguir:

- **Recuperação de Informação:** O módulo de Recuperação de Informação tem a finalidade de realizar a pesquisa em todos os documentos inseridos na base através dos metadados (e.g. titulo, autor, resumo, etc.) que o descrevem. Esta resposta à consulta é retornada ao usuário em ordem de relevância, provendo ao usuário fácil acesso às informações desejáveis.

- **Visualização de Documentos:** Após o usuário realizar a pesquisa no Clio-i, os registros são listados, com os metadados de cada documento. A partir daí, o usuário pode visualizar um documento específico de sua escolha, através do Visualizador de Documentos. O Visualizador de Documentos do Clio-i trabalha com arquivos do tipo áudio, imagem, vídeo e texto, dando a opção do usuário manipular o documento de maneira rápida e simples.

- **Clio-i Data Provider:** Conforme citamos no capítulo anterior, um Provedor de Dados OAI é um sistema que exporta os dados da sua base seguindo o protocolo OAI-PMH. Desta maneira, esse módulo exporta os dados de um repositório qualquer de acordo com a extensão do OAI-PMH que definimos (trataremos com detalhes na seção 4.3.3). Teremos então um sistema que fornece os serviços de uma Biblioteca Digital e que, agregado a ele, possua o Clio-i Data Provider para exportar a sua base de dados para futura coleta de um Provedor de Serviços qualquer.

- **Clio-i Service Provider:** Um Provedor de Serviços OAI realiza a coleta de diferentes repositórios, conforme detalhamos no capítulo 3. O Clio-i Service Provider realiza tanto

a coleta desses repositórios de acordo com o protocolo padrão do OAI (versão 2.0), quanto na versão estendida que nós adotamos. Além disso, os dados armazenados da coleta de outros repositórios são disponibilizados através dos serviços já mencionados e que podem ser coletados através do Clio-i Data Provider.

A partir desta explicação inicial, explicaremos com detalhes a base de dados utilizada no Clio-i, além de seus quatro módulos principais.

4.3.1 Base de Dados

A fim de atender a uma grande demanda de usuários, a base de documentos do Clio-i abriga diversas mídias, mais especificamente do tipo áudio, texto, imagem e vídeo. Os documentos que irão formar o acervo da Biblioteca Digital devem ser manualmente armazenados em uma base de documentos através do módulo de administração. Essa base de dados será posteriormente utilizada para recuperação de documentos de acordo com as consultas dos usuários finais.

Cada documento digital contido no sistema deve ser caracterizado por um conjunto de atributos descritores, os chamados metadados (e.g. autor, título, resumo, URL do documento, etc.). A base que armazena os metadados também faz uma referência ao documento eletrônico que fica localizado na base de documentos.

Desta forma, podemos especificar a base de dados do sistema em dois componentes: a Base de Metadados e a Base de Documentos.

Base de Metadados

O sistema Clio-i usa como base de seu conjunto de metadados os campos descritores definidos no padrão Dublin Core, visto com detalhes no capítulo 2. Na tabela 4.1 encontramos a relação entre os metadados do Dublin Core e a sua descrição específica para o Clio-i.

Tabela 4.1 Metadados descritivos dos documentos

Elemento Dublin Core	**Descrição**
Title	Título do documento.
Creator	Autor do documento
Description	Resumo do documento
Type	A natureza ou gênero do conteúdo. Descreve categorias gerais de documentos.(e.g., inconografia, cordéis, etc.).
Coverage	Onde o documento foi produzido / impresso
Publisher	A entidade responsável por tornar o documento disponível na sua forma atual.
Date	A data da produção / impressão do documento
Subject	As palavras-chaves que identificam o assunto documento
Contributor	Entidade responsável por contribuições ao conteúdo do recurso (e.g., tradutor, revisor, ilustrador, etc.)
Source	Documento do qual o presente documento foi derivado
Language	Principal idioma do conteúdo do documento
Format	Manifestação física ou digital do documento (e.g., jpeg, tiff, mp3)
Rights	Copyright, direitos do autor, propriedade intelectual
Relation	Referência para um documento relacionado.
Identifier	Uma referência não ambígua que identifica o documento (e.g., ID, URL).

Além dos metadados do Dublin Core, cada documento é associado ainda a informações específicas dos objetos digitais como resolução, duração (no caso de áudio e vídeo) e seqüência de páginas (para texto e imagem).

4.3.1.1 Base de Documentos

A base de documentos armazena os objetos digitais relacionados a cada documento (e.g. imagens JPG associadas a um documento). Essa base possui quatro divisões correspondentes aos tipos de mídia suportados: áudio, vídeo, imagem e texto. Cada documento é identificado por um código que o referencia na base de metadados.

4.3.1.2 Modelagem Entidade-Relacionamento

Os metadados do Clio-i e as informações dos objetos digitais serão armazenadas em um Banco de Dados Relacional. Um dos primeiros passos para a construção de qualquer Base Relacional é a construção de seu Modelo Conceitual [EN98]. Dentre estes tipos de modelagens, uma amplamente utilizada nos dias de hoje é a Modelagem Entidade-Relacionamento, modelo baseado na percepção do mundo real, que consiste em um conjunto de objetos básicos chamados entidades e nos relacionamentos entre esses objetos [SKS98].

Para facilitar o projeto do Banco de Dados do Clio-i, criamos uma Modelagem E-R representando o relacionamento entre os documentos e seus arquivos digitais, mostrados na figura 4.4.

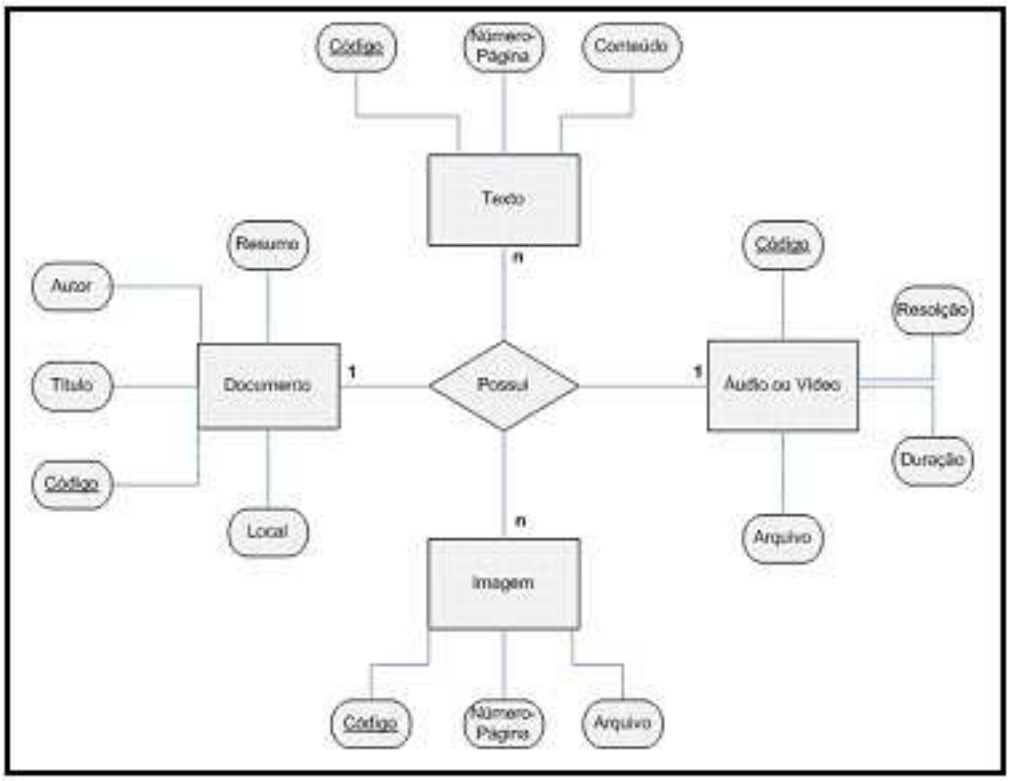

Figura 4.4 Modelagem Entidade-Relacionamento.

Nela, podemos perceber a entidade Documento, que recebe todos os atributos definidos na tabela 4.1. Cada documento, além de seus metadados descritivos, pode possuir um dos quatro tipos de mídias suportadas pelo Clio-i e cada mídia possui os seus atributos específicos. Ainda de acordo com a figura 4.4, um documento se relaciona apenas com um arquivo do tipo áudio ou vídeo, ou com vários objetos do tipo texto e imagem, que correspondem às páginas do documento.

4.3.2 Recuperação de Informação

Recuperação de informação é a representação, armazenamento, organização e acesso aos dados contidos em uma base de dados. A representação e organização dos dados devem prover fácil acesso ao usuário às informações que ao mesmo interessa [BYRN99]. A simples recuperação de dados, inseridos dentro do contexto de recuperação de informação, consiste em determinar que documento de uma coleção contém as palavras-chave da consulta de um usuário. Freqüentemente, isto não satisfaz o usuário, não retornando a informação realmente desejada. Assim, para que o usuário possa ter acesso aos dados contidos nas bases do Clio-i na forma mais relevante possível, foi projetado um módulo de Recuperação de Informação para o sistema.

Este módulo realiza a pesquisa nos metadados do documento (e.g. título, autor, resumo, etc.), retornando para o usuário um conjunto de registros, ordenados de acordo com a relevância das palavras-chave requisitadas. Há duas maneiras para utilizar este módulo no Clio-i. A

primeira delas é a pesquisa simples, onde é disponibilizado apenas um campo para o usuário digitar as palavras-chave desejadas, tornando a recuperação um trabalho simples e eficiente. Uma alternativa é a realização da pesquisa avançada, onde o usuário pode filtrar sua consulta, informando um formato de documento, coleção e/ou idioma específico. Além do mais, essa pesquisa conta com operadores booleanos (e.g. AND, OR, NOT), especializando ainda mais o trabalho do pesquisador.

4.3.3 Visualizador de Documentos

Como dito anteriormente, após a realização da consulta, o usuário pode escolher um dos registros retornados e visualizar o documento em questão, através deste módulo. Dentro do visualizador de documentos, algumas funcionalidades estão disponíveis, a fim de aumentar significativamente a acessibilidade e a usabilidade dos documentos digitais [CBG+05]. Todas essas funcionalidades são encontradas na tabela 4.2.

Tabela 4.2 Principais funcionalidades do Visualizador de Documentos

Suporte a Acervo Multimídia	O sistema aceita arquivos de texto, áudio, vídeo e imagem, descrito por um conjunto adequado de metadados.
Manipulador de documentos	O usuário pode manipular os documentos para uma melhor visualização (i.e., aumentar, diminuir, negativizar, girar, clarear escurecer e desfazer as manipulações).
Sistema colaborativo	O usuário pode inserir comentários sobre os documentos que serão compartilhados com outros usuários.
Opção de download	O usuário pode realizar download do documento sendo visualizado. Se o documento for uma imagem ou um texto, o sistema gera um arquivo PDF.
Menu de ajuda	O sistema provê um módulo de ajuda para os usuários que sentirem dificuldades na manipulação do documento.

Nas próximas seções apresentaremos com detalhes os dois módulos principais do Clio-i, responsáveis pela interoperabilidade de dados entre os repositórios digitais.

4.3.4 Clio-i Data Provider

No capítulo 3, explicamos com detalhes o funcionamento de um Provedor de Dados OAI. De fato, o projeto do Clio-i Data Provider segue a linha adota pelo *Open Archives*. A única exceção é que o módulo segue extensão do protocolo OAI-PMH adotada no presente trabalho e que será explicada mais a frente.

Alguns componentes se farão necessários para o perfeito funcionamento deste módulo. São eles:

- ***Parser:*** Tem a função de validar as requisições OAI.
- **Gerador XML:** Cria respostas XML com as informações dos metadados das requisições HTTP realizadas pelo Clio-i Service Provider ao Clio-i Data Provider.

- **Gerador de Erros:** Retorna, também codificado em XML, mensagens de erros de requisições mal-formadas (detalhes na seção 3.4).
- **Controle do Fluxo:** O controle de fluxo no Clio-i Data Provider é realizado para disponibilizarmos os metadados em porções específicas, conforme explicados no capítulo anterior.

Para ilustrarmos a arquitetura do Clio-i Data Provider, apresentamos a figura 4.5, que mostra os componentes e os seus relacionamentos:

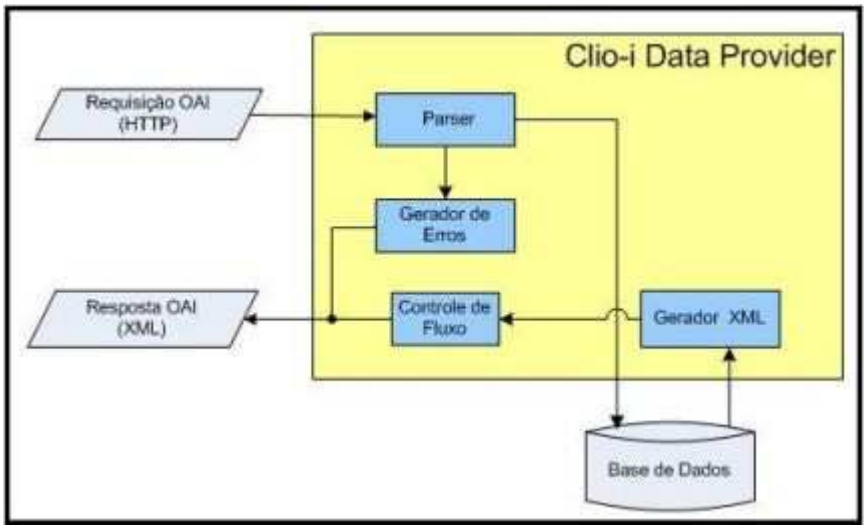

Figura 4.5 Arquitetura do Clio-i Data Provider.

O Clio-i Data Provider sempre recebe uma requisição HTTP como entrada. Alocado em algum servidor Web, o sistema recebe essa requisição e, havendo alguma má formação na mesma, uma mensagem de erro em formato XML é gerada e retornada para o requisitor. Caso contrário, através do *parser*, essa requisição HTTP é transformada em requisição SQL para a base de dados, que retorna ao sistema o resultado da consulta. Por fim, o gerador XML transforma o resultado da base de dados neste formato e o retorna para a saída do sistema, realizando o Controle de Fluxo dos dados exportados.

4.3.5 Clio-i Service Provider

Como já informado, um Provedor de Dados OAI realiza coleta dos metadados de repositórios espalhados pela Web, oferecendo algum serviço para usuários finais. O Clio-i Service Provider deve realizar a coleta tanto dos Provedores de Dados que estejam de acordo com a versão 2.0 do OAI-PMH (versão padrão), quanto àqueles provedores que estejam de acordo com a extensão do protocolo proposto nesse trabalho (neste caso, através da coleta dos metadados de um Clio-i Data Provider). Já os serviços que devem ser oferecidos serão realizados pelos módulos definidos na arquitetura do Clio-i (Recuperação de Informação e Visualizador de Documentos), explicados anteriormente.

Um Clio-i Service Provider deve possuir alguns componentes na sua arquitetura para o seu pleno funcionamento:

- **textitHTTP Request:** É um componente específico que envia requisições HTTP (através dos *havesters*) para os repositórios OAI a fim de coletar seus metadados.
- **Controle do Fluxo:** Para que possamos realizar o controle de fluxo no Clio-i Service Provider, é preciso que ele também seja implementado. O componente deve enviar novas requisições para o repositório requisitado, sempre que precisar de mais resultados.
- **Mecanismos de Atualização/Inserção:** É provável que alguns metadados que foram coletados há certo tempo, tenham sido modificados em seus repositórios originais. Desta forma, as informações no Provedor de Serviço estão desatualizadas e precisando de modificações. Esse componente deve permitir as duas formas de modificações na base de dados, conforme explicado no capítulo 2: a mais simples de todas é simplesmente apagar todos os metadados antigos antes de cada coleta nesse em um repositório específico (rematerialização da visão) ou uma alternativa é apagar e inserir novamente apenas os metadados atualizados no repositório (manutenção incremental).
- ***Parser XML:*** O Parser XML analisa a resposta dos Provedores de Dados às requisições para coleta e transforma a informação codificada em XML para a estrutura interna da base de dados do Clio-i.

Na figura 4.6, ilustramos a arquitetura do Clio-i Service Provider com todos os seus componentes.

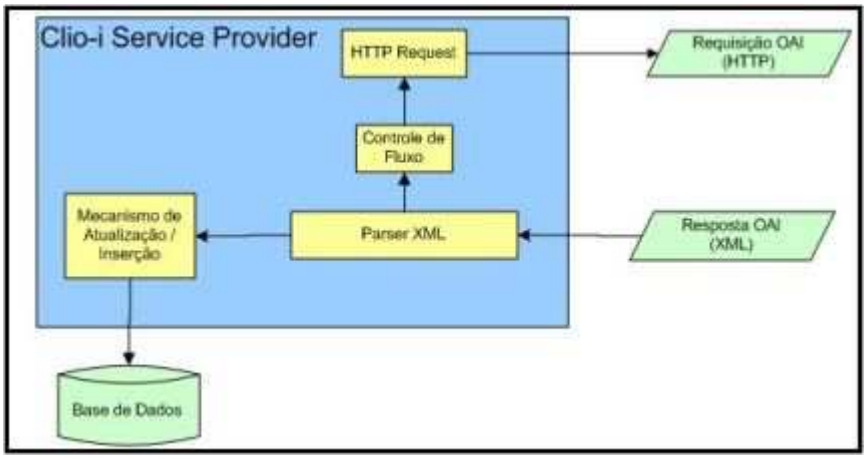

Figura 4.6 Arquitetura do Clio-i Service Provider.

Primeiramente um administrador do sistema realiza requisições a diferentes Provedores de Dados previamente cadastrados, através do componente *HTTP Request*. A resposta a essa requisição, em formato XML, é repassada então para o *parser XML*, que a codifica e repassa para o Mecanismo de Atualização/Inserção. Este componente insere os dados na base através

de comandos SQL. Caso o repositório requerido implemente o controle de fluxo, mais um ciclo de coleta é realizado automaticamente, até que os todos os dados requeridos estejam inseridos na base de dados.

4.4 Extensão do Protocolo OAI-PMH

No capítulo 3 discorremos sobre o protocolo para interoperabilidade entre repositórios digitais, o OAI-PMH. Na ocasião, detalhamos que as coletas são realizadas pelos Provedores de Serviços (especificamente os *havesters*), sobre os Provedores de Dados através de requisições HTTP. Esta requisição é composta de um dos seis verbos do protocolo, acompanhados dos seus parâmetros (ver tabela 3.3).

No mesmo capítulo 3, explicamos a facilidade e eficiência no uso do protocolo, visto que o mesmo trabalha a partir de tecnologias já bastante difundidas (e.g.: HTTP, Dublin Core, XML). Entretanto, no levantamento dos requisitos do Clio-i, identificamos algumas carências do OAI-PMH e, a partir delas, tentamos estender o protocolo para atender as nossas necessidades.

Inicialmente, imaginemos um Provedor de Dados hipotético que possua milhões de registros armazenados sobre documentos científicos. Se quisermos restringir essa consulta, poderíamos utilizar os parâmetros *until*, *from* e *set* (explicados no capítulo anterior) com os verbos *ListRecords* ou *ListIdentifiers*. Entretanto, digamos que se pretenda coletar apenas registros que citem o termo "Redes Neurais"em um dos seus metadados. A versão mais recente do protocolo OAI-PMH - a 2.0 - não permite isso, ou seja, hoje teríamos que coletar os milhões de dados previamente para depois realizar essa seleção.

Desta maneira, encontramos uma solução bastante prática para o problema. Nos dois verbos que retornam a lista dos registros - ListRecords e ListIdentifiers - adicionamos mais um argumento, a que chamamos de *query*. Este novo argumento receberia um conjunto de termos que serão pesquisáveis dentro dos metadados do repositório (e.g. title, description, creator, etc.), já citados anteriormente.

Agora, para coletarmos os registros sobre Redes Neurais deste repositório hipotético, poderíamos realizar a seguinte requisição:

http://repositorio.com/oai/index.php?verb=ListRecords
&metadataPrefix=oai_dc&query="Redes Neurais"

Além desta lacuna, nos deparamos com uma necessidade básica, mas que o protocolo OAI-PMH não cobre diretamente: a de saber a quantidade de registros em um repositório. O que o protocolo sugere é adicionarmos ao elemento *resumptionToken* (explicado com detalhes no capítulo anterior) um parâmetro chamado *completeListSize*.

Entretanto, poucos Provedores de Dados adicionam essa informação tão relevante. Por exemplo, a *Library for digital documents at the university of Oslo*[5], Provedor de Dados filiado ao *Open Archives*, demonstra essa informação. Vamos supor que realizamos a seguinte requisição a este repositório:

[5]Library for digital documents at the university of Oslo - http://wo.uio.no/

http://wo.uio.no/as/WebObjects/theses.woa/wa/oai?verb=ListRecords
&metadataPrefix=oai_dc

Como vimos no capítulo anterior, a informação deverá ser retornada através de um XML de acordo com os padrões do OAI-PMH. Uma parte do documento XML correspondente a essa requisição específica, como pode ser visto na figura 4.7.

Figura 4.7 Exemplo do parâmetro *completListSize*.

Como podemos perceber no destaque da figura 4.7, é retornado o tamanho total de registros desta requisição. Entretanto, verificando a importância da notificação de quantos registros são retornados em uma certa requisição, decidimos que esse dado deveria constar obrigatoriamente nas respostas em XML e da forma mais clara possível.

Desta maneira, foi decidido criar um outro verbo para suprir essa necessidade do protocolo, chamado de *GetSize*. Este sétimo verbo do protocolo poderia retornar a quantidade exata de registros de um repositório, ou informar o tamanho do repositório através de condições de refinamento da consulta, utilizando os parâmetros *until*, *from*, *set* (explicados no capítulo anterior) e *query* (o novo parâmetro criado).

Assim, uma requisição com este verbo, utilizando o repositório científico hipotético citado anteriormente, poderíamos realizar o seguinte tipo de requisição:

http://repositorio.com/oai/index.php?verb=GetSize
&from=2005-30-01&query='Redes Neurais'

Com esta requisição, pretendemos saber o numero de registros que foram inseridos na base a partir do dia 30 de janeiro de 2005 e que possua em algum dos seus metadados o termo Redes Neurais. A resposta em XML da requisição em questão pode ser verificada na figura 4.8.

Figura 4.8 Resposta à requisição do verbo GetRecord.

No capítulo anterior, mostramos a tabela 3.3, que se refere aos verbos e parâmetros encontrados no protocolo OAI-PMH. Como o Clio-i permite trabalharmos com um novo parâmetro e um novo verbo, agora apresentamos na tabela 4.3 uma reformulação do quadro anterior, agora de acordo com a extensão proposta.

Para finalizar, relatamos no capítulo anterior que os Provedores de Dados dificilmente expõem o recurso em si para a coleta, apenas os metadados. E quando assim o fazem, geralmente é apenas uma URL com alguma informação básica. Deve-se constatar, entretanto, que alguns repositórios enviam o caminho de um arquivo html ou pdf, por exemplo, do recurso por completo, mostrados na figura 4.9, do Provedor de Dados *Université du Québec*[6].

Figura 4.9 Exemplo da URL de um recurso completo.

Como podemos identificar no destaque da figura 4.9, o elemento *dc:identifier* contém a URL para o documento por completo. Por outro lado, alguns repositórios que disponibilizam o recurso, o colocam no elemento *dc:source*, não apresentando desta maneira uma uniformidade do protocolo na descrição dos seus recursos.

[6]Université du Québec - http://sdeir.uqac.ca/

Tabela 4.3 Os verbos e seus argumentos de acordo com a extensão do protocolo OAI-PMH

Verbo	**Descrição**	**Argumentos**
GetRecord	Recupera os metadados de um item individual de um repositório.	**identifier**. Obrigatório. Com ele, especificamos o identificador único (ver seção 3.2.1) do item de um repositório. **metadataPrefix**. Obrigatório. Especifica o padrão de metadados adotado que deve estar especificado no Provedor de Dados.
Identify	É usado para coletar informações sobre um repositório.	Não há argumentos.
ListRecords	Este verbo recupera os metadados de um repositório.	**from.** Opcional. Os dados coletados devem ser criados ou alterados a partir da data específica por este argumento. **until.** Opcional. Os dados coletados devem ser criados ou alterados até a data especificada pelo argumento. **metadataPrefix.** Já explicado anteriormente. **set.** Opcional. Especifica um conjunto, para o *havester* poder refinar a sua coleta. **resumptionToken.** Exclusivo. Argumento necessário quando os provedores utilizam o controle de fluxo na coleta dos metadados. **query.** Opcional. Termo que será pesquisado nos metadados do repositório
ListIdentifiers	É uma abreviação do ListRecords, que retorna apenas o *header* (ver seção 3.2.1) de um repositório.	**from.** **until.** **metadataPrefix.** **set.** **resumptionToken.** **query.**
ListMetadataFormats	Retorna os padrões de metadados utilizados em um repositório.	**identifier.** Opcional (apenas neste verbo). Retorna o padrão de metadados utilizado em um item específico.
ListSets	É utilizado para retornar a estrutura de um repositório, listando todos os conjuntos que compõe os metadados	**resumptionToken.**
GetSize	Finalidade de retornar a quantidade de registros em uma requisição	**from.** **until.** **metadataPrefix.** **set.** **query.**

O Clio-i foi projetado para operacionalizar com arquivos do tipo áudio, texto, vídeo e imagem. Quando falamos nessa operacionalização, estamos nos referindo à visualização do documento multimídia e a sua interoperabilidade. Assim, para atender ao intercâmbio desses tipos de dados, simplesmente acrescentamos um novo elemento dentre os metadados de um recurso, a qual chamaremos de *dc:clioidocument*. Desta maneira, além de não deixar mais dúvidas quanto ao metadado utilizado para a fonte do recurso, identificamos que o documento pertence a uma base ligada ao sistema Clio-i. Na descrição dos metadados de um recurso, este novo elemento poderá receber a URL de um vídeo ou áudio, das páginas de um arquivo do tipo imagem, ou o conteúdo das páginas de um arquivo do tipo texto.

4.5 Critérios de Desenvolvimento

Após os estudos realizados, definimos algumas características e funcionalidades que o software necessita para atender as expectativas de uma Biblioteca Digital completa. Basicamente, o sistema interage com três atores:

- **Administrador:** Coordena as atividades do Clio-i Data Provider e do Clio-i Service Provider.
- ***Havester:*** Trata-se de um robô pertencente ao Clio-i Service Provider, mas que coleta informações de Provedores de Dados OAI previamente cadastrados no Clio-i Service Provider.
- **Usuário Comum:** Interage com as bases do Clio-i Service Provider e do Clio-i Data Provider através dos serviços oferecidos pelo sistema.

Estes atores interagem com as funcionalidades do sistema em três cenários diferentes, que são justamente os três módulos principais do Clio-i explicados anteriormente.

Na próxima seção detalharemos cada funcionalidade apresentada nos diagramas acima.

4.5.1 Requisitos

A seguir apresentaremos os requisitos detalhadamente de cada módulo do Clio-i. Além disso, em cada módulo mostraremos uma figura representando os atores e as funcionalidades, apresentando uma visão em alto nível do sistema nos seus quatro módulos principais. Tal visão auxiliará na compreensão de cada cenário, e ajudará na identificação dos atores e dependências entre requisitos do sistema. Cada uma das elipses representa uma funcionalidade e as setas entre os atores e funcionalidades representam relações entre os mesmos.

Recuperação de Informação e Visualizador de Documentos

A figura 4.10 corresponde ao Diagrama de Casos de Uso dos módulos que oferecem os serviços aos usuários de uma Biblioteca Digital. Vale salientar que na figura adicionamos o Caso de Uso inserir mensagens no mural de recados.

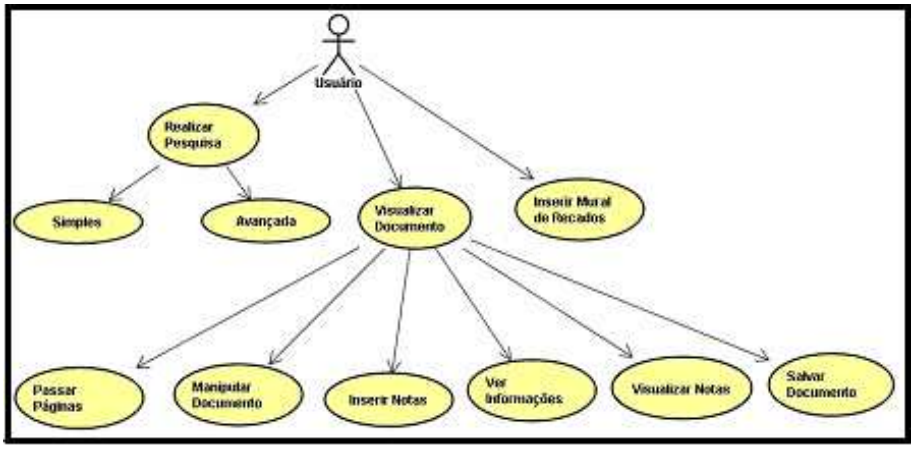

Figura 4.10 Diagrama de Casos de Uso dos serviços oferecidos.

A partir daí, descreveremos as principais funcionalidades dos casos de uso descritos, cujo ator em todas elas é o Usuário.

- **Pesquisar Documentos:** Todos os documentos inseridos na base devem ser pesquisáveis através dos metadados (e.g. titulo, autor, resumo, etc.) que o descrevem. O resultado dessa pesquisa deve listar os registros em ordem de relevância em relação às palavras-chave pesquisadas, conforme mencionado no início do capítulo. Como podemos perceber na figura 4.10, esta funcionalidade pode ser efetuada através de uma pesquisa simples ou de uma pesquisa avançada:
  Pesquisa Simples: É oferecido ao usuário apenas um campo para ele digitar as palavras-chaves. Após digitar os termos de seu interesse, basta clicar no botão pesquisar para efetivar a consulta.
  Pesquisa Avançada: São oferecidas ao usuário opções de refinamento da consulta. As opções são as seguintes: pesquisa com operadores booleanos (e.g. AND, OR e NOT). O usuário pode realizar também a consulta por documentos de uma mídia específica, por coleções onde os documentos estão agrupados ou por um idioma específico.

- **Visualizar Documento:** Depois de realizada a pesquisa, o usuário poderá clicar em algum link para abrir o documento em questão através de um Visualizador de Documentos. O módulo deve reproduzir documentos do tipo áudio, vídeo, texto ou imagem. Dentro deste módulo, algumas funcionalidades estão disponíveis. São elas:
  Manipulação do Documento: Essa funcionalidade foi adicionada para que os usuários possam se familiarizar com a manipulação de documentos reais. Desta maneira, o Visualizador de Documentos deve permitir que o usuário realize as seguintes manipulações no objeto digital: aumentar ou diminuir, inverter horizontalmente ou verticalmente, clarear ou escurecer e negativar. Vale salientar que muitas vezes o efeito de negativação é necessário para a melhor visualização de documentos com estado de conservação baixa.
  Inserir Notas: Um usuário pode pretender realizar algumas anotações sobre o documento (e.g. transcrição de documentos manuscritos, anotações de partes relevantes do

texto, dúvidas, etc.). Para isso, este módulo deve oferecer uma área para que os usuários possam realizar essas anotações sobre um documento específico. A mesma deve ser armazenada na base de dados do Clio-i e enviadas por e-mail para o usuário.
Visualizar Notas: Permite que usuários possam visualizar todas as anotações realizadas sobre um documento específico.
Salvar Documentos: Se algum usuário pretende possuir esse documento digital (e não apenas acessá-lo pelo Visualizador), o sistema deve permitir que assim o faça. Para isso, deve conter uma opção para o usuário realizar o download do recurso. No caso de documentos do tipo texto e imagem, o sistema deve disponibilizar ao usuário arquivos do tipo PDF. É de fundamental importância que esse arquivo PDF seja gerado no momento do pedido do usuário, para reduzir os curtos de armazenamento dos dados.
Ver Informações: O visualizador deve não só disponibilizar o documento em sim, mas possuir uma opção para que o usuário possa saber todas as informações a respeito do mesmo (e.g.: titulo, autor, resumo, etc.)

Agora, detalharemos as funcionalidades dos dois módulos principais do sistema, responsáveis pela interoperabilidade do Clio-i.

Clio-i Data Provider

Na figura 4.11 apresentamos o Diagrama de Casos de Uso referente ao Clio-i Data Provider. Nele, podemos observar os dois atores do módulo e todas as suas funcionalidades correspondentes.

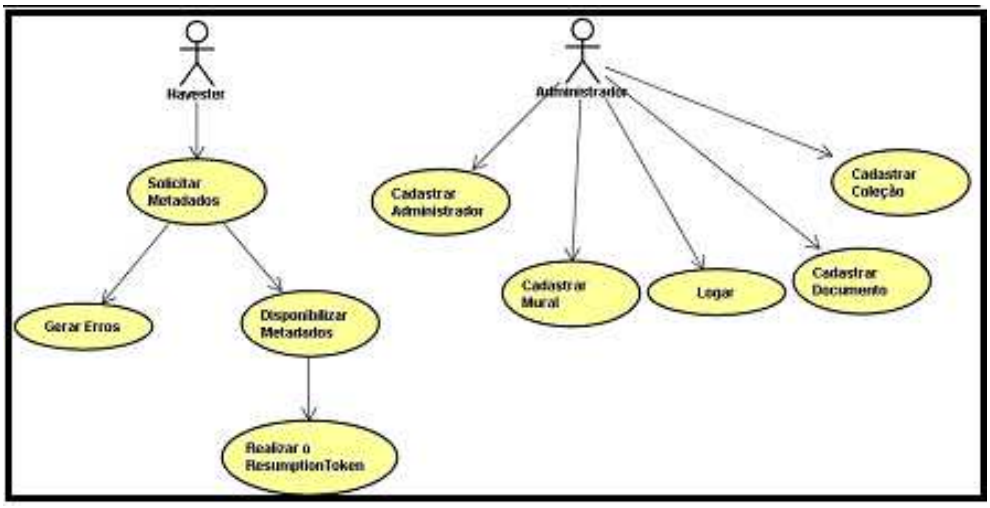

Figura 4.11 Diagrama de Casos de Uso do Clio-i Data Provider.

Apresentamos a seguir as principais funcionalidades do Administrador do Clio-i Data Provider.

- **Cadastrar Documento:** Permite que o administrador do Sistema realize a inserção, alteração, consulta e exclusão dos documentos na base de dados.

- **Cadastrar Coleções:** Cada recurso em uma base de dados do Clio-i pertence a uma coleção específica, tornando a disponibilização e a sua coleta mais organizadas. Desta maneira, essa funcionalidade permite que administradores insiram, alterem, excluam e consultem coleções da base de dados.

As funcionalidades seguintes são atribuídas ao ator *Havester*, coletor de dados presente nos Provedores de Serviços OAI.

- **Solicitar Metadados:** Um *Havester* faz uma requisição HTTP ao Clio-i Data Provider, que retorna essa requisição ao mesmo através de informações codificadas em XML. Essa requisição deve obedecer à extensão do protocolo OAI-PMH, obedecendo a seus sete verbos e todos os seus parâmetros, definida pelo projeto e que será explicada ainda neste capítulo. Nesta funcionalidade são incluídas outras duas:
  Gerar de Erros: Se uma solicitação HTTP não estiver de acordo com o protocolo, um erro é gerado e retornado para que o *Havester*. Os erros que podem ser gerados foram explicados no capítulo anterior.
  Receber Metadados: Se a solicitação for bem formatada, de acordo com o que foi detalhado no capítulo 3, o Clio-i Data Provider disponibiliza os metadados formatados de acordo com a extensão do protocolo OAI-PMH. Essa disponibilização deve implementar o controle de fluxo através do resumptionToken, que é uma alternativa para não enviar todos os recursos de uma só vez, os disponibilizando em porções de tamanhos fixos. Todos os detalhes deste controle de fluxo foram explicados no capítulo anterior.

Clio-i Service Provider

Por fim, apresentamos na figura 4.12 o Diagrama de Casos de Uso correspondente ao Clio-i Service Provider, mostrando o seu ator e funcionalidades.

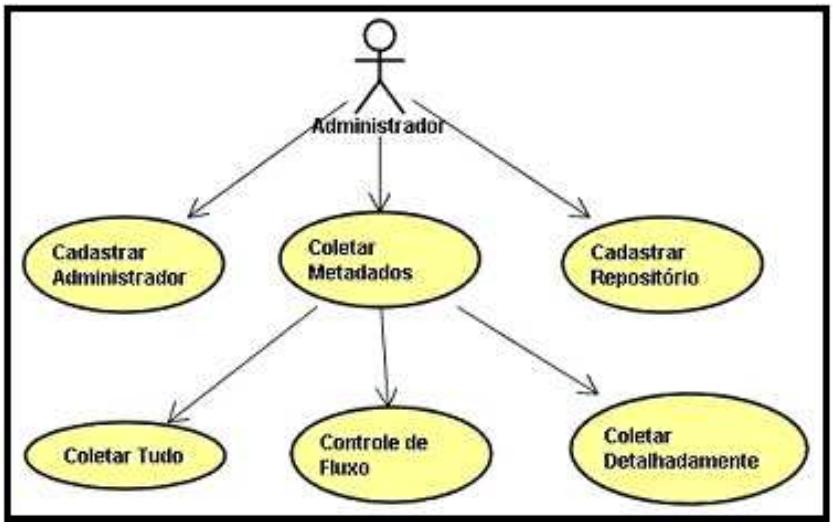

Figura 4.12 Diagrama de Casos de Uso do Clio-i Service Provider.

O Clio-i Service Provider tem o objetivo principal de realizar a coleta dos metadados de repositórios OAI. Conforme constatamos na figura 4.12 ele só possui um ator (o Administrador) e as principais funcionalidades exercidas pelo mesmo são detalhadas a seguir:

- **Cadastrar Repositório:** Os Provedores de Dados OAI que serão coletados pelo Clio-i Service Provider devem ser cadastrados previamente no sistema. Desta maneira, esta funcionalidade permite que o administrador do sistema consiga inserir, alterar, excluir e consultar um repositório existente na Internet.

- **Coletar Metadados:** Uma das principais de todo o Clio-i, essa funcionalidade tem o objetivo de coletar os metadados (e os recursos, quando disponibilizados) dos repositórios cadastrados através de um *Havester*. Vale ressaltar que esse *Havester* deve coletar de repositórios OAI que estejam de acordo com o protocolo OAI-PMH versão 2.0 ou de acordo com a extensão do OAI-PMH, sem a necessidade de saber previamente qual das duas versões é utilizada pelo Provedor de Dados cadastrado. Esta funcionalidade inclui outras duas, que são:
  Coletar todos os metadados: Realiza a coleta de todos os metadados de um Provedor de Dados qualquer de uma só vez. Se já existirem os metadados coletados, os mesmos serão apagados e os novos serão incluídos em seu lugar.
  Coletar com detalhes: Ao invés de coletar todos os metadados de uma só vez, podem-se aplicar refinamentos nessa coleta. Assim, o administrador pode restringir os metadados através da data de inclusão no Provedor de Dados e pelo conjunto a qual ele pertence. Ressaltamos aqui que a data da última coleta deve ser armazenada, para que o administrador possa coletar apenas os dados mais recentes de um repositório.

Ambas as coletas devem suportar o controle de fluxo, caso os repositórios a implementem.

4.6 Considerações Finais

Após um estudo detalhado das necessidades de uma Biblioteca Digital, realizamos a proposta do sistema Clio-i, detalhados neste capítulo. Até chegar à versão proposta, o sistema atravessou vários protótipos, sendo adicionados a ele sugestões e críticas de instituições, pesquisadores e usuários.

Além da vantagem de apresentar o documento por completo e o módulo de Recuperação de Informação, o Clio-i permite a interoperabilidade entre bases digitais, utilizando uma extensão do já difundido protocolo OAI-PMH. Essa extensão fez-se necessário devido a pequenas necessidades que achávamos necessários para o perfeito funcionamento do integrador de dados.

Para construir o sistema, diversas tecnologias foram utilizadas, metodologias de desenvolvimento foram aplicadas e testes realizados. Desta maneira, os detalhes de implementação do Clio-i serão apresentados no próximo capítulo.

CAPÍTULO 5

Protótipo e Testes

Pense duas vezes antes de agir. Aja duas vezes antes de pensar. Bifurque, diferencie, simule, salve como.

—EDUARDO LOUREIRO JR.

No capítulo 4, apresentamos a arquitetura geral do Clio-i, sistema para gerenciamento de Bibliotecas Digitais que, dentre diversas características, permite a interoperabilidade entre repositórios digitais. Este capítulo tem a finalidade de descrever detalhes de implementação do protótipo atual do Clio-i, relacionando a metodologia utilizada na sua construção, requisitos para funcionamento, estrutura completa da base de dados e os módulos de processamento (Recuperação de Informação, Visualizador de Documentos, Clio-i Data Provider e Clio-i Service Provider). Apresentamos ainda testes funcionais e de usabilidade realizados com o protótipo implementado. Os estudos de caso onde o Clio-i foi aplicado serão apresentados no capítulo 6.

5.1 Metodologia de Construção

Relatamos no capítulo 4, acerca da metodologia aplicada no desenvolvimento do Clio-i, onde identificamos uma construção evolutiva do software. De fato, diversas versões do sistema foram apresentadas e avaliadas até chegarmos à versão atual, apresentada neste trabalho. E esse sempre foi o objetivo ao construirmos o sistema, utilizando a abordagem de prototipação de software como forma de validação de requisitos rapidamente.

A prototipação de software é utilizada como uma técnica de análise e redução de riscos, cujos erros e omissões dos requisitos estão entre os mais comuns [Som01]. Assim, aliamos uma abordagem evolucionária do sistema, em que uma versão inicial foi produzida rapidamente - no caso, a primeira versão do projeto Ultramar (ver seção 4.2) - e incrementalmente foi modificada para produzir a versão atual.

A partir deste desenvolvimento interativo e incremental, utilizamos a prototipação evolucionária, que se baseia na idéia de desenvolver uma implementação inicial, expondo-as aos comentários dos usuários e aperfeiçoando-a ao longo de alguns estágios, até que um sistema adequado tenha sido desenvolvido [Som01], conforme ilustra a figura 5.1.

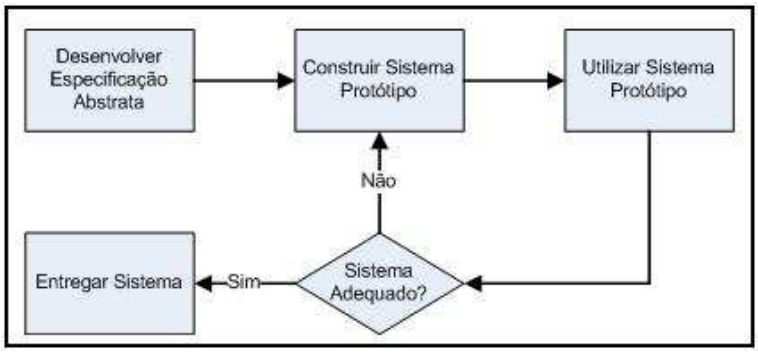

Figura 5.1 Ciclo da prototipação evolucionária.

Até o desenvolvimento do Clio-i nos dias atuais, passamos pelo desenvolvimento de alguns protótipos: Versão inicial do Ultramar, Biblioteca Digital para Documentos Históricos, Biblioteca Digital Multimídia e Sistema Clio-i. Desta maneira, passamos por quatro ciclos de prototipação evolutiva até atingir o estado atual, como mostrado na figura 5.2.

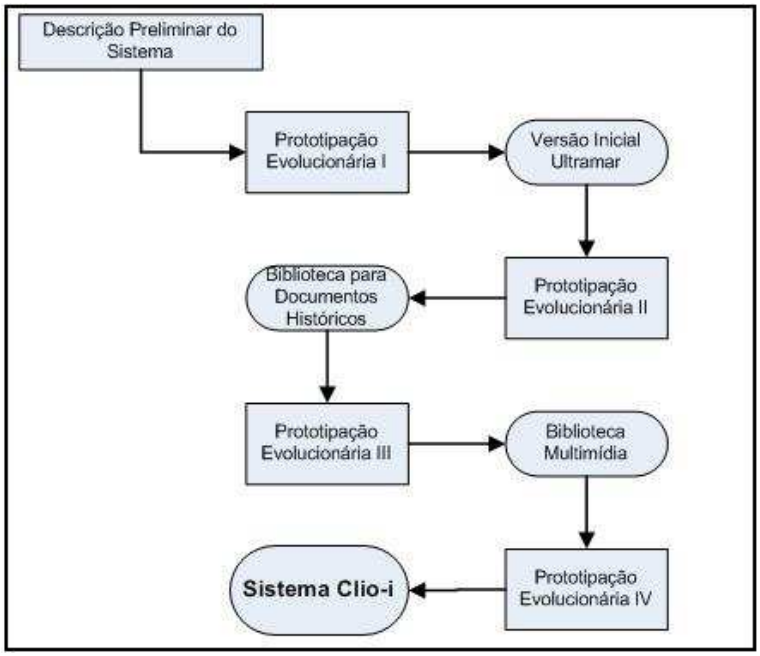

Figura 5.2 Ciclos de protótipos até o Clio-i.

5.2 Características Gerais

O desenvolvimento do protótipo buscou seguir critérios de qualidade de Engenharia de Software. Como visto no capítulo 4, o Clio-i possui uma arquitetura modular, buscando favorecer

a extensibilidade e adaptabilidade do sistema. O sistema é disponibilizado com código aberto com o objetivo de permitir modificações/adaptações mais complexas do software.

Os documentos da base são armazenados em formatos difundidos (e.g., texto ASCII, JPG, PNG, WAV, MP3, etc.), não oferecendo problemas para visualização fora do ambiente do sistema - caso o usuário faça o download do documento para consultas futuras. O sistema foi implementado em PHP [PHPa], uma linguagem adequada para criação de páginas dinâmicas na Web. O sistema Clio-i pode ser utilizado facilmente em computadores que possuam servidores Web (e.g., Apache Server). Especificamente, todos os testes do protótipo foram realizados no seguinte ambiente: Apache Server (versão 2.2.2) com suporte à linguagem PHP (versão 5.2) e servidor de bando de dados MySQL (versão 4.1).

Primeiramente, é uma obrigatoriedade que o sistema trabalhe a partir da versão 5 do PHP, pois apenas a partir dela que a linguagem oferece suporte à biblioteca SimpleXML [PHPb], que será vital para a aplicação do integrador de dados, explicado mais a frente. Estamos utilizando a versão 5.2 do PHP, pois somente a partir dela que a linguagem oferece suporte "nativo"ao Apache 2.2.2.

Por se tratar de um sistema Web, tivemos a preocupação do mesmo funcionar em diversos navegadores, como o Internet Explorer e o Mozila FireFox.

5.3 Estrutura da Base de Dados

O Sistema Gerenciador de Banco de Dados (SGBD) escolhido para o armazenamento dos metadados foi o MySQL [MySb], o banco de dados *open-source* mais popular do mercado. Esse SGBD fornece recursos simplificados e apropriados para as suas aplicações, tendo um custo extremamente reduzido. Possui ferramentas que atendem a maioria das exigências de uma aplicação de base de dados corporativa, fornecendo uma arquitetura extremamente rápida e de fácil utilização. Desta forma, destacamos as seguintes características para a escolha do MySQL: confiabilidade e bom desempenho, facilidade de utilização e distribuição, recursos e suporte para as mais diversas plataformas.

No capítulo anterior, mostramos o primeiro passo da construção de nossa base de dados, com um resumo da Modelagem Entidade-Relacionamento do sistema (ver figura 4.4). Apresentamos aqui a modelagem completa da base de dados do Clio-i, apresentada na figura 5.3.

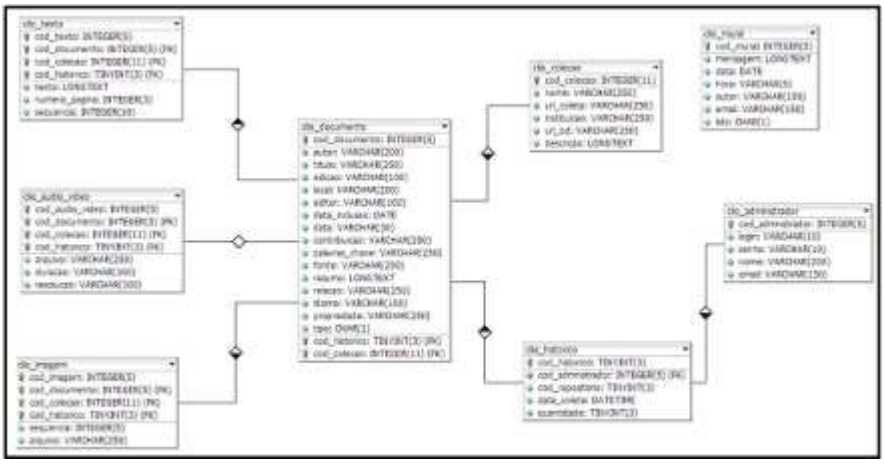

Figura 5.3 Modelo Relacional da base de dados.

A tabela *clio_documento* refere-se à entidade documento, mostrada na figura 4.4 e, por conseguinte, as tabelas *clio_audio_video*, *clio_imagem* e *clio_texto* referem-se aos objetos digitais associados aos documentos. A tabela *clio_nota* armazena todas as notas enviadas por usuários a respeito de um documento específico, assim como a *clio_mural* as mensagens no mural de recados postadas por visitantes do sistema. A tabela *clio_colecao* tem a função de armazenar as coleções específicas de cada documento. No caso de um Clio-i Service Provider, a tabela serve para armazenar informações dos Provedores de Dados que serão coletados, caso contrário, guarda informações dos conjuntos para organizar os documentos de um Clio-i Data Provider.

5.4 Módulos de Processamento

A seguir apresentamos os detalhes de implementação dos módulos de processamento do sistema: Visualizador de Documentos, Recuperação de Informação, o Clio-i Data Provider e o Clio-i Service Provider.

5.4.1 Visualizador de Documentos

Como já foi explicado anteriormente, o este módulo de documentos foi criado para atender a uma demanda de visualização do documento pela WEB com uma boa usabilidade, para que usuário possa manipulá-lo de forma simples e rápida.

Para tal, implementamos o visualizador de documentos utilizando DHTML (*Dynamic HTML*) [GS03]. O DHTML vem ganhando popularidade como um método interativo de visualização de informação. Muitas implementações desta linguagem vem sendo estudadas e o interesse vem crescendo para o design de páginas na Internet. A tecnologia foi empregada tanto nas operações sobre os documentos, quanto na construção das janelas móveis resultando num módulo satisfatório em termos de usabilidade, velocidade e design.

Na figura 5.4, apresentamos o módulo manipulando um documento do tipo imagem.

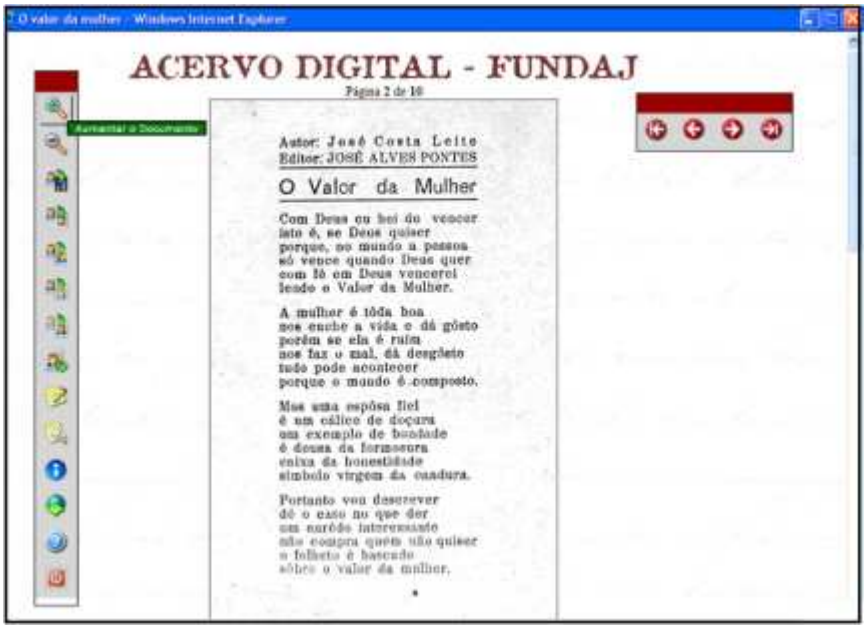

Figura 5.4 Visualizador de Documentos.

Neste exemplo, demonstramos a página de um documento depositado no Acervo Digital da Fundação Joaquim Nabuco, que utiliza o sistema Clio-i (detalhes do estudo de caso no capítulo seguinte). Nela, encontramos duas barras de ferramentas. No canto esquerdo, temos a primeira barra com botões que ativam os seguintes efeitos sobre o documento: aumentar, diminuir, negativar, inverter verticalmente, inverter horizontalmente, clarear, escurecer, e restaurar o documento original. Ainda há outras opções nessa barra, como inserir notas sobre o documento, visualizar as notas, visualizar as informações sobre o documento, realizar o download em PDF, um menu de ajuda sobre o Visualizador e sair do módulo. Ainda na figura 5.4, no canto superior direito encontramos a segunda barra de ferramentas, que tem a função de realizar a navegação sobre as páginas do documento. Vale salientar que essas barras de ferramentas são móveis, permitindo que o usuário as posicione no local que achar mais adequado.

No caso da figura 5.4, apresentamos um exemplo de visualização de um documento do tipo imagem. Contudo, o visualizador, como dito anteriormente, suporta ainda arquivos do tipo áudio, vídeo e texto. A figura 5.5 mostra um exemplo de um vídeo sendo reproduzido no Visualizador de Documentos. Ele funciona com a ajuda de um algum *player* que esteja instalado na máquina do usuário (e.g. Windows Media Player, QuickTime, etc.).

Figura 5.5 Reprodução de um Vídeo no Visualizador de Documentos.

Por fim, apresentamos mais uma funcionalidade do Visualizador de Documentos: a de inserir notas. Clicando nesta opção na barra de ferramentas, uma janela (também móvel) aparecerá para o usuário preencher nome, e-mail e a nota em si. A composição desta nota é auxiliada por um mini-editor de texto, como pode ser percebido na figura 5.6. Nele, o usuário pode deixar o texto em negrito, inserir algum link, imagens, alterar a formatação do texto, etc.

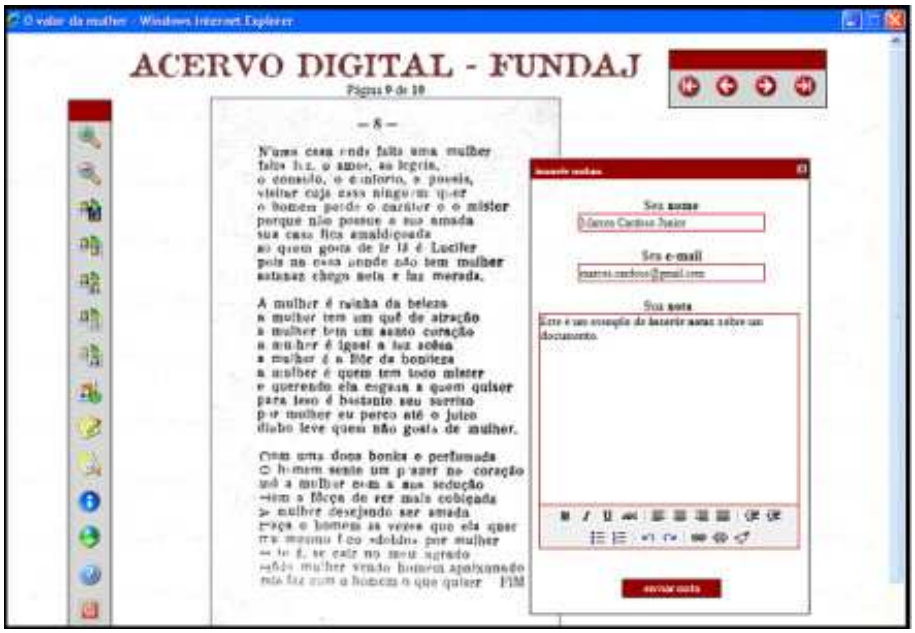

Figura 5.6 Inserindo notas sobre o Documento.

Vale salientar que todas as operações dentro do Visualizador de Documentos são efetuadas sem a necessidade de se carregar a página Web. Isso se dá pela facilidade que a tecnologia DHTML - aplicada a estas operações - oferece.

5.4.2 Recuperação da Informação

No capítulo 4, apresentamos uma pequena definição de Recuperação de Informação e a necessidade de um módulo desse porte para auxiliar nas pesquisas dos documentos. Dentro deste contexto, nosso objetivo era construir um módulo de Recuperação da Informação simples e eficiente. Então, nesse módulo resolvemos utilizar um recurso do banco de dados MySQL, o chamado *MySQL Full-Text Search* (versão 3.23.23) [MySa]. Dentre as características desse recurso temos:

- Disponibiliza funções usadas para casamento de palavras-chave e expressões de busca booleana nos campos do banco de dados.
- Ordena os resultados da busca usando um critério de relevância.
- Bastante eficiente, apresentando um tempo de resposta aceitável.

Uma consulta básica utilizando o *MySQL Full-Text Search* se assemelha com uma consulta SQL comum. A diferença é a utilização das funções *MATCH()* e *AGAINST()*. Para realizar um consulta dentro da tabela clio_documento com as palavras-chave "Brasil"e "Cultura", a seguinte consulta deveria ser utilizada:

SELECT * FROM clio_documento WHERE MATCH (titulo,resumo)
AGAINST ('Brasil Cultura')

Como podemos perceber, nesse exemplo, os campos de metadados pesquisados são definidos na função *MATCH()* e as palavras-chave na função *AGAINST()*. Quando estas funções são utilizadas na cláusula *WHERE*, conforme mostrado, os resultados são retornados e ordenados por ordem de relevância. Essa relevância é medida através da freqüência com que os termos de consulta são encontrados nos campos do documento, ponderados pela freqüência com que os termos estão presentes na base de documentos.

Com o *MySQL Full-Text Search*, também podemos utilizar operações booleanas, com a ajuda do modificador *IN BOOLEAN MODE*. A seguir, temos um exemplo deste tipo de consulta.

SELECT * FROM documentos WHERE MATCH (titulo,resumo)
AGAINST ('+Brasil - Cultura' IN BOOLEAN MODE)

Esta consulta recupera todos os documentos que contenham o termo "Brasil", mas que não possuam a palavra "Cultura". O *IN BOOLEAN MODE* suporta outras operações, exemplificadas abaixo, que são argumentos da função *AGAINST*:

- **'Brasil Cultura'.** Encontra documentos que contenha pela menos uma destas palavras.
- **'+Brasil +Cultura'.** Encontra documentos que contenham exatamente ambas as palavras.
- **'Brasil*'.** Retorna documentos que contenham as palavras brasil, brasileira, brasileiro, brasileiros, etc.

- **'+Brasil Cultura'.** Encontra registros que contenham obrigatoriamente a palavra Brasil, mas possuirá uma relevância maior se também possuir o termo Cultura.
- **' "Brasil Cultura"'.** Encontra registros que contenham exatamente o termo "Brasil Cultura".

Antes de enviarmos as palavras-chave ao processamento, realizamos o que chamamos de eliminação de *stop words* da consulta [BYRN99]. Para isso, foi implementada uma função auxiliar ao módulo, a fim de eliminar palavras que não possuem um valor semântico associado ao documento. Encontra-se nesse conjunto palavras muito freqüentes na base (que já são eliminadas pelo *MySQL Full-Text Search*) ou termos como artigos, preposições, conjunções, alguns advérbios e adjetivos. Assim, antes do processamento de casa consulta, o módulo de Recuperação da Informação do Clio-i realiza a eliminação das *stop words* para um melhor aproveitamento da pesquisa realizada pelo usuário.

Quanto ao tempo de resposta das pesquisas utilizando o *MySQL Full-Text Search*, o resultado foi bastante satisfatório. Todos os projetos onde o Clio-i foi utilizado, o tempo médio de resposta às consultas foi em torno de 2 segundos, inclusive para bases muito grandes, como a do primeiro estudo de caso que será visto no capítulo seguinte.

5.4.3 Clio-i Data Provider

Apresentamos no capítulo anterior, os quatro componentes que fazem parte do Clio-i Data Provider. Para este módulo, um conjunto de funções foi implementado e as principais estão representadas na figura 5.7, que mostra um fluxograma do funcionamento deste módulo. Cada função está ligada ao seu módulo, como podemos perceber na legenda da figura.

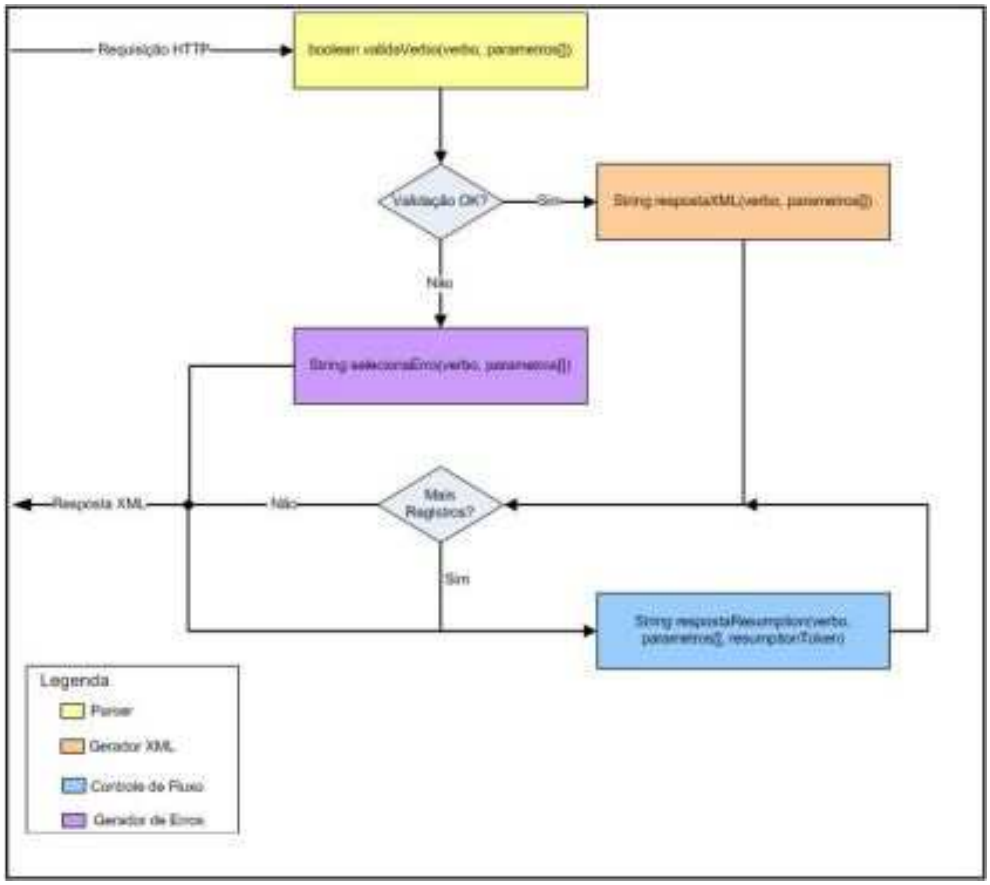

Figura 5.7 Principais funções e componentes do Clio-i Data Provider.

Tecemos alguns comentários a respeito da implementação de cada módulos nas próximas seções.

Parser

Para este componente, foi implementado um conjunto de funções com a finalidade de validar as requisições OAI. Digamos que o parser receba a seguinte URL de requisição:

http://www.urlbasica.com/oai/index.php?parametro1=valor1
&parametro2=valor2&parametro3=valor3

O parser analisa todos os parâmetros da URL requisitada, realizando essa validação em dois passos:

- **Análise do verbo:** É analisado se foi passado algum verbo como parâmetro e se esse verbo consta em um dos sete definidos no capítulo anterior.
- **Análise dos argumentos:** Após o primeiro passo realizado com sucesso, é verificado se os argumentos do verbo também estão de acordo a extensão do protocolo adotada.

A principal função deste componente é a *validarVerbo*, mostrado na figura 5.7. Recebendo o verbo passado na requisição e um conjunto de parâmetros, ele realiza os dois passos descritos acima para então decidir para que componente à informação será passada.

Gerador XML

Caso a requisição HTTP tenha sido aprovada pelo *parser*, a função *respostaXML* retorna uma string no formato XML com todas as informações requisitadas, levando em conta o verbo solicitado e seus argumentos.

Paralelo a isso, imagine que um recurso do tipo texto possua duzentas páginas de conteúdo digital. Dessa maneira, quando este registro for exportado em XML, haverá duzentos elementos *dc:cliodocument* para relacionar as páginas desse documento. Com a possibilidade real dos metadados exportados do Clio-i serem de grandes tamanhos, tomamos a decisão de retornar apenas 100 registros de cada vez na sua exportação.

Ou seja, se a resposta da solicitação possuir mais do que esse limite estabelecido, o componente Controle de Fluxo do Clio-i Data Provider é acionado.

Controle de Fluxo

Como dito anteriormente, caso a requisição validada possua mais de 100 registros em sua resposta, esse componente é ativado através da função *respostaResumption*. Essa função retorna a mesma resposta em XML que é gerada pelo componente anterior. A diferença é que a função do Controle de Fluxo recebe um parâmetro a mais, que é justamente o *resumptioToken*, responsável por realizar a paginação dos resultados em XML das requisições.

Gerador de erros

Este componente é chamado caso o *parser* retorne falso para a validação da requisição HTTP. Através da função *selecionaErro*, é verificado que erro foi cometido e sua resposta em formato XML é apresentada.

5.4.4 Clio-i Service Provider

Assim como fizemos no módulo passado, a implementação do Clio-i Service Provider foi desenvolvida com base nos componentes descritos no capítulo anterior, cada um com algumas funções. As principais funções e seus respectivos componentes são ilustrados no fluxograma da figura 5.8.

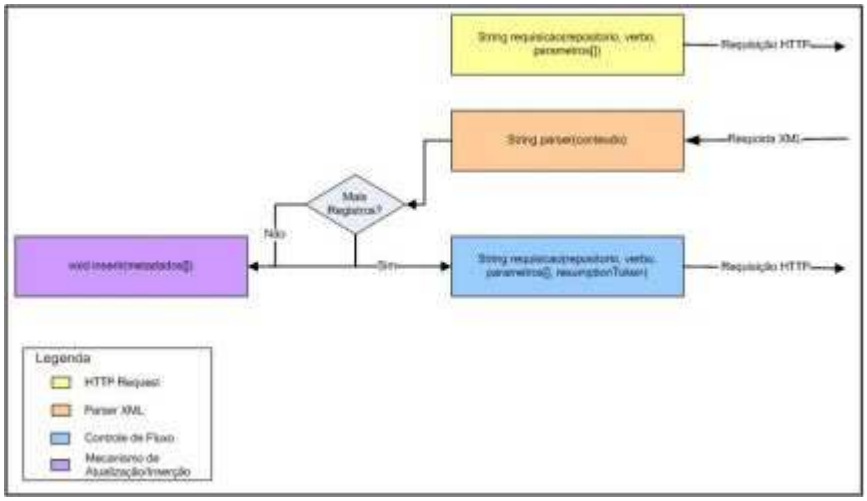

Figura 5.8 Principais e componentes do Clio-i Service Provider.

HTTP Request

Passo inicial para a coleta dos dados, este componente realiza a requisição para um Provedor de Dados previamente cadastrado no sistema. O *HTTP Request* possui a função requisicao, onde passamos como parâmetro o repositório escolhido, o verbo e seus possíveis argumentos. Dessa maneira, a função seleciona a URL Básica para coleta do repositório e monta a requisição com base nos parâmetros passados.

Parser XML

Após a requisição, este componente recebe o conteúdo XML através da função parser, e faz uma chamada ao componente Mecanismos de Inserção/Atualização para inserir os dados no banco. O *Parser XML* ainda verifica se há mais requisições a serem feitas. Caso ainda haja porções de registros a serem coletadas do repositório selecionado, o componente Controle de Fluxo é chamado.

Controle do Fluxo

Após a confirmação que existem registros a serem coletados, a função requisicao é acionada, essa do componente Controle de Fluxo. Aliás, essa função é bem parecida à homônima encontrada no *HTTP Request*, tendo como única diferença o parâmetro adicional *resumptionToken*.

Mecanismos de Atualização/Inserção

Recebe do *parser XML* os valores necessários para o componente realizar a atualização da base de dados do Clio-i Service Provider. Através da função inserir, os metadados coletados são inseridos ou atualizados, caso o registro coletado já tenha sido armazenado na base alguma outra vez. Ressaltamos que essa atualização é necessária para que a base do Clio-i Service

Provider esteja de acordo com a base do Provedor de Dados em questão.

SimpleXML

Para a realização da coleta de diferentes Provedores de Dados OAI utilizamos uma biblioteca nativa do PHP chamada SimpleXML [PHPb]. Esta biblioteca visa integrar, de maneira simples, XML no PHP, trabalhando numa estrutura de objetos. Resumidamente, a figura 5.9 nos mostra um trecho de código em que estamos realizando a coleta do Provedor de Dados Analytical Sciences Digital Library[1] .

```
1  <?php
2
3  $conteudo = new SimpleXMLElement("http://www.asdlib.org/oai/oai.php?verb=ListRecords&metadataPrefix=oai_dc");
4
5  foreach($conteudo->ListRecords->record as $registro ) {
6
7     $row = simplexml_load_string($registro->asXML());
8
9     $row->registerXPathNamespace('dc', 'http://purl.org/dc/elements/1.1/');
10
11    $title = $row->xpath('//dc:title');
12
13    $creator = $row->xpath('//dc:creator');
14
15    $subject = $row->xpath('//dc:subject');
16
17    mysql_query("INSERT INTO clio_documento(titulo, autor, resumo) VALUES ('$title', '$creator', '$subject')");
18
19 }
20
21 ?>
```

Figura 5.9 Trecho de código utilizando o SimpleXML.

Primeiramente, a variável *$conteudo* recebe uma instância de *SimpleXMLElement*, referente ao nó raiz do XML passado, no caso, a requisição para a coleta (linha 3). Após isso, é realizado um laço, percorrendo cada registro do arquivo XML passado (linha 5). A linha 7, por sua vez, se faz necessário para carregarmos cada registro desses como um XML diferente, por exigência da própria biblioteca. Prosseguindo, informamos no capítulo 2 que os elementos Dublin Core dentro de um XML são especificados de uma maneira peculiar (e.g. dc:title, dc:creator, etc.). Desta maneira, na linha 9 encontramos o método *registerXPathNamespace*, cuja finalidade é criar um prefixo para validar o nome de um elemento Dublin Core. Após isso, coletamos três metadados de um registro (title, creator e subject), colocando-os em variáveis, como mostra as linhas 11, 13 e 15. Finalmente, ocorre a inserção dos metadados na base de dados, através do código SQL apresentado na linha 17.

5.5 Testes

Nesta seção iremos tratar dos dois testes realizados no Clio-i. O primeiro deles foi o Teste Funcional, que verifica se todos os requisitos do sistema foram corretamente implementados. O segundo teste realizado trata-se do Teste de Usabilidade, responsável por medir a complexidade no uso do Clio-i pelos usuários.

[1] Analytical Sciences Digital Library - http://www.asdlib.org/

5.5.1 Teste Funcional

Em todo processo de desenvolvimento de software, existe uma grande possibilidade de injeção de falhas humanas. Os custos associados a essas falhas justificam um processo de testes cuidadoso e bem planejado [Pre92].

A aplicação de testes funcionais, ou testes de caixa preta, é uma abordagem nas quais os testes são derivados da especificação do programa, onde o testador determina o comportamento do sistema através de suas entradas e saídas relacionadas [Som01].

O método de caixa preta foi utilizado para testar os requisitos do Clio-i, através dos casos de testes estabelecidos e descritos mais à frente. Os requisitos que serão testados estão descritos na tabela 5.1.

Tabela 5.1 Requisitos do Clio-i a serem testados

Requisito	**Módulo**
Pesquisar Documento	Recuperação de Informação
Visualizar Documento	Visualização de Documentos
Cadastrar Coleção	Clio-i Data Provider
Cadastrar Documento	Clio-i Data Provider
Cadastrar Repositório	Clio-i Service Provider
Coletar Metadados	Clio-i Service Provider

Cada requisito do Clio-i a ser testado é associado a um Caso de Teste, que contém a descrição do caso, descreve as pré-condições de execução do teste, os passos específicos do teste a ser executado e os resultados esperados após a sua execução (pós-condições). Nas tabelas seguintes, apresentamos todos os Casos de Testes de cada requisito.

Tabela 5.2 Caso de Teste Pesquisar Documento

[CT01] Pesquisar Documento	
Descrição	Esse caso de teste tem como objetivo validar a pesquisa de um documento no sistema. Deve ser validada a busca simples e a avançada.
Passos	Realize 10 pesquisas simples, aumentando o número de termos em cada uma delas (a primeira pesquisa com um termo, a segunda com dois e assim sucessivamente). Realize 10 pesquisas avançadas, aumentando o número de termos em cada uma delas (a primeira pesquisa com um termo, a segunda com dois e assim sucessivamente). Nestas pesquisas avançadas, devem-se testar todos os campos booleanos (OR, AND e NOT), em coleções específicas e com uma mídia escolhida.
Pré-condição	A base de dados deve estar com pelo menos 100 registros de documentos.
Pós-condição	Consulta retornada para o usuário, em ordem de relevância dos termos pesquisados e atendendo aos refinamentos da pesquisa avançada. Os termos que foram inseridos na pesquisa devem vir destacados no texto. O tempo de resposta não pode ultrapassar 5 segundos.

Tabela 5.3 Caso de Teste Visualizar Documento

[CT02] Visualizar Documento	
Descrição	Esse caso de teste tem como objetivo validar a visualização de documentos do tipo texto, imagem vídeo e áudio no sistema.
Passos	Abra um documento do tipo imagem. Realize todas as manipulações possíveis (aumentar, diminuir, negativar, clarear, escurecer, inverter horizontalmente, inverter verticalmente e restaurar para o formato original). Realize o download em PDF do documento, insira notas, visualize as notas disponíveis e cheque as informações sobre um documento. Por fim, saia do visualizador através da opção da barra de ferramentas. Realize as mesmas operações para um documento do tipo áudio, outro do tipo vídeo e outro do tipo texto.
Pré-condição	A base de dados deve conter pelo menos um documento de cada mídia armazenado.
Pós-condição	Todas as manipulações e operações sobre o documento devem ser atendidas com sucesso.

Tabela 5.4 Caso de Teste Cadastrar Coleção

[CT03] Cadastrar Coleção	
Descrição	Esse caso de teste tem como objetivo validar a inserção, alteração, consulta e exclusão das coleções em uma base de dados do Clio-i.
Passos	Insira uma coleção qualquer na base. Nos campos referentes à coleção, não tentar inserir sem preencher campos obrigatórios. Alterar a coleção cadastrada. Nos campos referentes à coleção, tentar inserir sem preencher campos obrigatórios. Consultar a coleção alterada através de termos referentes a ela. Excluir a coleção.
Pré-condição	Estar logado no sistema de administração.
Pós-condição	Coleção inserida com sucesso, verificando na base de dados o novo registro. Coleção alterada com sucesso, verificando na base de dados a alteração do registro. Retorno da coleção através dos termos solicitados na consulta. Exclusão da coleção inserida, verificando na base de dados a ausência da mesma.

Tabela 5.5 Caso de Teste Cadastrar Documento

[CT04] Cadastrar Documento	
Descrição	Esse caso de teste tem como objetivo validar a inserção, alteração, consulta e exclusão dos documentos em uma base de dados do Clio-i.
Passos	Insira quatro documentos na base, cada um com uma mídia diferente. Nos campos referentes ao documento, tentar inserir sem preencher campos obrigatórios. Alterar o documento cadastrado. Nos campos referentes ao documento, não preencher campos obrigatórios. Consultar o documento alterado através de termos referentes a ela. Excluir o documento.
Pré-condição	Estar logado no sistema de administração.
Pós-condição	Documento inserido com sucesso, verificando na base de dados o novo registro. Documento alterado com sucesso, verificando na base de dados a alteração do registro. Retorno do documento através dos termos solicitados na consulta. Exclusão do documento inserido, verificando na base de dados a ausência do mesmo.

Tabela 5.6 Caso de Teste Cadastrar Repositório

[CT05]Cadastrar Repositório	
Descrição	Esse caso de teste tem como objetivo validar a inserção, alteração, consulta e exclusão dos repositórios que terão seus metadados coletados em uma base de dados.
Passos	Insira um repositório qualquer na base. Nos campos referentes ao repositório, tentar inserir sem preencher campos obrigatórios. Alterar o repositório cadastrado. Nos campos referentes ao repositório, tentar inserir sem preencher campos obrigatórios. Consultar o repositório alterado através de termos referentes a ele. Excluir o repositório.
Pré-condição	Estar logado no sistema de administração.
Pós-condição	Repositório inserido com sucesso, verificando na base de dados o novo registro. Repositório alterado com sucesso, verificando na base de dados a alteração do registro. Retorno do repositório através dos termos solicitados na consulta. Exclusão do repositório inserido, verificando na base de dados a ausência do mesmo.

Tabela 5.7 Caso de Teste Coletar Metadados

[CT06] Coletar Metadados	
Descrição	Esse caso de teste tem como objetivo validar a coleta dos metadados de um Provedor de Dados previamente cadastrado.
Passos	Selecione um Provedor de Dados cadastrado. Primeiro, realize a coleta de todos os metadados do repositório. Selecione outro Provedor de Dados. Realize a coleta, especificando conjunto a ser coletado, data de início e data fim.
Pré-condição	Estar logado no sistema de administração. Algum repositório previamente cadastrado.
Pós-condição	Metadados coletados inseridos na base. Realizar verificação dos metadados coletados com especificação dos metadados do Provedor de Dados.

Todos os testes foram realizados com sucesso e as imperfeições do sistema identificadas foram tratadas. O relatório completo do resultado do Teste Funcional pode ser visto no Apêndice A - Relatório de Avaliação dos Testes Funcionais.

5.5.2 Teste de Usabilidade

A usabilidade de um sistema é relacionada à eficácia e eficiência da interface diante do usuário e pela reação do usuário diante da interface, descrevendo a qualidade da interação entre homem-máquina [Fer02]. Qualidade esta associada por alguns princípios no uso de um sistema [HH94]:

- Facilidade de aprendizado e memorização de tarefas no caso de uso intermitente.
- Produtividade dos usuários na execução de tarefas.
- Prevenção, visando a redução de erros por parte do usuário.
- Satisfação subjetiva do usuário.

O teste de usabilidade utilizado para o sistema Clio-i envolveu alguns tipos de tarefas buscando encontrar problemas de usabilidade e fazer recomendações no sentido de eliminar os problemas e melhorar a usabilidade do produto.

Vale ressaltar que todo o Teste de Usabilidade aplicado ao Clio-i é baseado no trabalho proposto em [Fer02], incluindo metodologia e documentos de apoio ao teste.

Primeiramente, tivemos que elaborar um Plano de Teste de Usabilidade (ver Apêndice B), documento que descreve o propósito do teste, a descrição dos problemas que queremos solucionar, os perfis dos usuários utilizados, a metodologia aplicada, o papel do avaliador neste teste de usabilidade e as medidas de avaliação que serão adotadas ao fim do teste.

Segundo [Fer02], um número de quatro a cinco participantes é capaz de identificar a maioria dos problemas de Usabilidade de um sistema. Assim, decidimos reunir quatro participantes para realizar o Teste de Usabilidade proposto.

Decidimos escolher participantes com diferentes perfis e experiências no uso de sistemas Web, para assim, avaliar o Clio-i sob o ponto de vista de usuários distintos. Dois deles, são experientes quanto à utilização da Internet, principalmente sobre sistemas de buscas e cadastros. Um outro participante possui uma experiência baixa em relação a esses tipos de sistemas e a sua escolha se deveu ao fato da possibilidade de adaptarmos o sistema para usuários inexperientes. Por fim, o último participante possui uma experiência razoável relacionados a sistemas como o Clio-i.

Durante os testes, foi apresentada a cada participante uma lista de tarefas a serem executadas dentro do Clio-i, documento este que pode ser encontrado no Apêndice C - Lista de Tarefas para o Teste de Usabilidade. Com a ajuda de uma planilha, os seguintes dados foram anotados para concretizarmos os resultados finais: tempo gasto para a execução da tarefa, número de erros, se conseguiu ou não realizar a tarefa e algumas observações que possam ser importantes. Após o teste, além dos dados coletados durante o procedimento, outros dados são necessários para se medir a usabilidade do sistema. Assim, foi oferecido a cada usuário um questionário de avaliação do Clio-i, disponível no Apêndice D - Questionário de Avaliação do Sistema pelo

Participante. Após isso, em uma conversa aparentemente informal, foram levantados alguns questionamentos a serem discutidos sobre o Clio-i.

Finalizado os testes e, conseqüentemente as coletas dos dados, realizamos um estudo sobre as reais necessidades de Usabilidade do Clio-i. Em termos gerais, o software foi bem aceito pelos participantes, que não tiveram dificuldades em realizar a maioria das tarefas solicitadas. Entretanto, pequenas modificações foram sugeridas e todos os detalhes do resultado dos testes serão vistos a seguir.

Resultados

Os resultados dos testes de usabilidade no Clio-i foram baseados nos seguintes itens: tempo de execução das tarefas, número de erros na execução de cada tarefa e resposta do questionário de avaliação pelos participantes.

Na tabela 5.8, apresentamos as medidas coletadas na execução das vinte e duas tarefas realizadas no teste de usabilidade. As mesmas foram comparadas com valores previamente estabelecidos como: pior nível aceitável para a execução das tarefas e nível alvo pretendido para a execução de cada tarefa. A tabela também apresenta o valor médio das medidas coletadas.

Tabela 5.8 Tempo de Execução das Tarefas (em segundos)

Tarefas	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Pior Nível	180	120	120	60	60	120	120	60	120	180	60
Nível Alvo	100	70	70	35	35	70	70	35	70	100	35
Usuário 1	28	39	19	14	10	51	27	18	80	60	18
Usuário 2	119	132	116	79	37	97	79	59	239	113	59
Usuário 3	41	34	28	32	19	48	24	31	98	68	13
Usuário 4	98	58	75	43	49	112	65	40	130	101	22
Média	71,5	65,7	59,5	42	28,7	77	48,7	37	136,7	85,5	28
Tarefas	**12**	**12**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**
Pior Nível	300	180	240	180	60	300	240	180	300	180	180
Nível Alvo	180	100	140	100	35	180	140	100	180	100	100
Usuário 1	122	99	214	110	10	57	219	67	199	45	113
Usuário 2	190	171	235	197	54	169	211	187	340	89	177
Usuário 3	78	86	76	47	12	68	136	50	184	37	43
Usuário 4	176	81	132	102	38	99	189	84	221	58	119
Média	141,5	109,2	164,2	114	28,5	98,2	188,7	97	236	37,2	113

Na tabela 5.9 é apresentado o número de erros de cada participante nas vinte e duas tarefas estabelecidas. Semelhante à tabela 5.8, nesta apresentamos uma comparação com os valores pior de nível aceitável e nível alvo pretendido, calculando ainda a média de erros por tarefa.

Tabela 5.9 Número de erros por tarefas

Tarefas	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Pior Nível	1	2	2	1	2	3	3	1	2	4	1
Nível Alvo	0	1	1	0	1	1	1	0	1	2	0
Usuário 1	0	0	0	0	2	1	1	0	3	1	0
Usuário 2	1	2	0	2	1	4	3	0	11	3	1
Usuário 3	0	0	0	1	0	0	2	0	5	0	0
Usuário 4	1	1	1	3	2	6	0	0	9	2	0
Média	0,50	0,75	0,25	1,50	1,25	2,75	1,50	0	7,00	1,50	0,25
Tarefas	**12**	**12**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**
Pior Nível	3	2	3	4	1	4	4	3	4	3	3
Nível Alvo	1	1	1	2	0	2	2	1	2	1	1
Usuário 1	0	1	2	1	0	0	2	0	2	0	0
Usuário 2	2	4	5	7	0	2	8	3	3	0	1
Usuário 3	0	0	1	2	0	0	5	1	0	0	0
Usuário 4	1	2	2	4	0	0	4	1	4	0	0
Média	0,75	1,75	2,5	3	0	0,50	4,75	1,25	2,25	0	0,25

Por fim, apresentamos a tabela 5.10 com as respostas dos participantes (Pt.) do teste ao questionário de avaliação do sistema onde as respostas são oferecidas ao participante em uma escala 0 a 5.

Tabela 5.10 Caso de Teste Coletar Metadados

	Pt. 1	**Pt. 2**	**Pt. 3**	**Pt. 4**	**Média**
Facilidade de utilização	5	4	4	5	4,5
Organização das informações	4	4	5	4	4,25
Nomenclatura utilizada nas telas (nome de comandos, títulos, campos, etc.)	3	3	2	3	2,75
Mensagens do sistema	5	4	5	4	4,5
Assimilação das informações	4	3	4	5	4
Satisfação na realização dos testes	3	3	3	4	3,25

Como podemos perceber nos dados coletados, em termos gerais, a aceitação quanto à usabilidade do sistema foi boa. Mesmo usuários com pouca experiência conseguiram realizar todas as tarefas, na maioria das vezes, abaixo do pior nível aceitável de tempo estipulado.

Entretanto, há de se destacar alguns aspectos. Podemos perceber, principalmente na tabela 5.10, que as nomenclaturas utilizadas no sistema não satisfizeram os participantes. De fato, isso foi observado principalmente no sistema de administração para coleta de dados. Na figura 5.10 encontramos a área em que os participantes tiveram grandes dificuldades para cumprirem as suas tarefas.

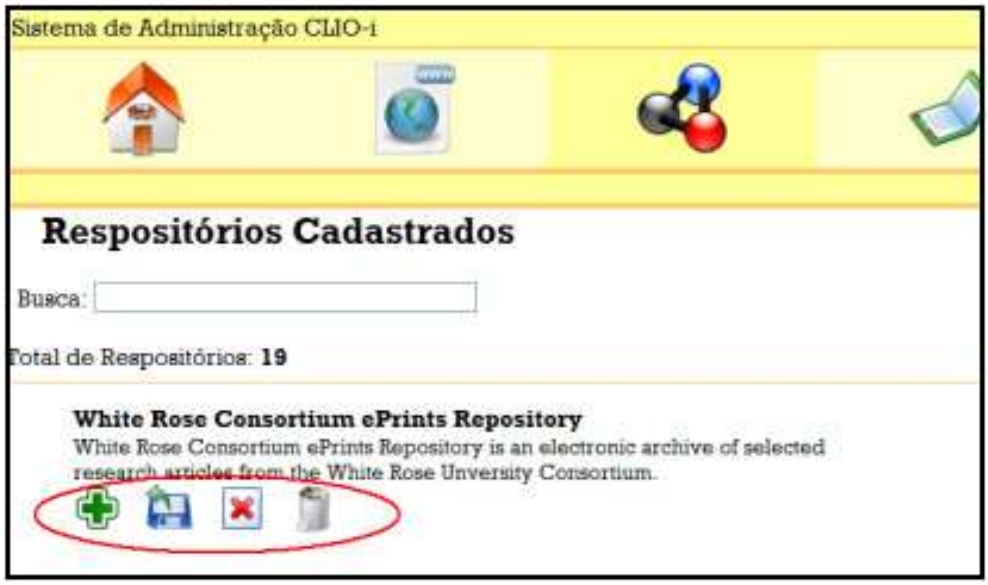

Figura 5.10 Área para coleta dos dados.

Os ícones apresentados na figura 5.10 para a realização de tarefas como coletar todos os metadados e coletar dados com detalhes está longe de intuitivo, tendo que muitas vezes o usuário testar o ícone para acertar a atividade. A mudança realizada foi simples, colocando abaixo da figura o que representa cada tarefa, textualmente.

Outro aspecto em que todos os participantes tiveram dificuldades foi na localização da pesquisa avançada. Pelo fato da opção da atividade apenas aparecer na página inicial (e com um texto pequeno), a duração de sua localização rendia um tempo considerável. A saída foi colocarmos a pesquisa avançada como uma opção principal do sistema. Assim, seja a atividade que o usuário esteja fazendo no Clio-i, a opção de pesquisa avançada é oferecida, conforme mostramos na figura 5.11.

Figura 5.11 Opção de Busca Avançada após o Teste de Usabilidade.

5.6 Considerações Finais

Neste capítulo realizamos diversas considerações sobre os detalhes de implementação do protótipo do Clio-i. Construído dentro da metodologia de prototipação, sobre um desenvolvimento interativo e incremental, o sistema passou por diversas versões, a fim de atender a uma demanda cada vez mais crescente nessa área.

Optou-se pelo uso da linguagem PHP na construção do sistema, por ser uma das mais adequadas para criação de páginas dinâmicas na Web. O Banco de Dados escolhido foi o MySQL, que fornece recursos simplificados e apropriados para as suas aplicações, com um custo extremamente reduzido, se comparado com os demais.

Para validar a proposta de um sistema robusto e que atendesse a demanda nessa área, foi realizado dois tipos de testes no sistema: Testes Funcionais e Testes de Usabilidade. Com os testes, conseguimos obter respostas a algumas falhas no sistema para o seu melhor aproveitamento.

Para concretizar o trabalho, foram realizados dois estudos de caso sobre o Clio-i, que serão detalhados no capítulo seguinte.

CAPÍTULO 6

Estudos de Caso

Em vez de utilizar os computadores para ensinar, como se tentava fazer há vinte anos, eles nos servem para a validação dos conhecimentos. O trabalho nobre é deixado para os homens.

—PIERRE LÉVY & MICHAEL AUTHIER

Nos dois capítulos anteriores, detalhamos o sistema Clio-i. No capítulo 4, tecemos considerações sobre suas funcionalidades, extensões do protocolo OAI-PMH e arquitetura do sistema completo, bem como de seus módulos. No capítulo 5, relatamos sobre o protótipo do Clio-i, descrevendo seus detalhes de implementação. Ainda realizamos testes funcionais e de usabilidade para verificar limitações e aprimorarmos o sistema.

Neste capítulo, apresentaremos dois estudos de caso realizados com o Clio-i. No primeiro deles, reunimos bases de metadados provenientes de diversos Provedores de Dados registrados no *Open Archives Initiative*. O segundo estudo de caso foi realizado com o objetivo de avaliar em um ambiente real as funcionalidades empregadas no Clio-i, como manipulação de documentos de diversas mídias e utilização da extensão do protocolo OAI-PMH.

6.1 Estudo de Caso 1: Integrador de Repositórios Científicos

O primeiro estudo de caso apresentado corresponde ao Integrador de Repositórios Científicos[1] . A proposta desse projeto era reunir em uma única base diversos metadados correspondentes a arquivos eletrônicos científicos. O objetivo principal desse estudo de caso era fazer que um Clio-i Service Provider possuísse uma base com milhares de documentos vindos de diversos lugares para que algumas funcionalidades do Clio-i fossem mais bem avaliadas na prática.

O desempenho das consultas do módulo de Recuperação de Informação, por exemplo, poderia ser avaliado com uma base maior. Outro módulo que poderia ser aina melhor avaliado seria o Clio-i Service Provider, que coletaria informações de um número maior de provedores de dados.

Para realizamos o estudo de caso, foi realizada uma consulta à página dos Provedores de Dados oficiais da OAI para selecionarmos alguns repositórios científicos. Na tabela 6.1, apresentamos os Provedores de Dados selecionados (total de 19 provedores), destacando o nome do provedor, uma pequena descrição (retirada do site do provedore) e a URL básica para coleta dos dados.

Após essa seleção manual de repositórios científicos, deu-se continuidade à construção da base de metadados, com o cadastro de cada um dos 19 repositórios selecionados e a coleta

[1] Integrador de Repositórios Científicos - http://www.liber.ufpe.br/clioi

Tabela 6.1 Provedores de Dados OAI selecionados

Provedor de Dados OAI	Descrição	URL básica para coleta
White Rose Consortium ePrints Repository	Repositório que armazena arquivos eletrônicos de artigos selecionados da White Rose Unversity Consortium.	http://sherpa.leeds.ac.uk/perl/oai2
Université du Québec à Chicoutimi - Documentation régionale	Serviço de documentação sobre estudos históricos da Université du Québec. Reúne documentos sobre geografia, história, ciência política, entre outros.	http://sdeir.uqac.ca/metadata/oai.asp
Universitätsbibliothek Marburg	A biblioteca da Universidade de Marburg possibilita a publicação de trabalhos científicos por seus alunos, professores e pesquisadores.	http://archiv.ub.uni-marburg.de/oai/oai2.php
Universität Karlsruhe: EVA - Elektronisches Volltextarchiv	Esta é uma Biblioteca da Universidade de Karlsruhe que armazena documentos científicos da entidade.	http://www.ubka.uni-karlsruhe.de/oai/eva/oai2.php
University of Wuerzburg, GERMANY, OPUS	OPUS é um serviço de publicação on-line que oferece aos seus membros a oportunidade de publicar seus documentos na Internet para a posterior busca de usuários externos.	http://opus.bibliothek.uni-wuerzburg.de/oai/oai2.php
University of Wisconsin Digital Collections	A coleção da Wisconsin foi criada no verão de 2001 para prover recursos acadêmicos de qualidade vindos da biblioteca da Universidade.	http://oaidp.library.wisc.edu/oaicat/OAIHandler
University of Pittsburgh Electronic Thesis and Dissertation Archive	Reúne Teses e Dissertações em formato eletrônico, especificamente em arquivos PDF.	http://etd.library.pitt.edu/ETD-db/NDLTD-OAI2/oai.pl
University of Joensuu: electronic publications	Documentos eletrônicos da University of Joensuu. Inclui teses e publicações em formato PDF.	http://joypub.joensuu.fi/OAI/
Universidad Nacional de La Plata	Este serviço oferece acesso a teses, dissertações e diversos tipos de criações intelectuais sobre arte, tecnologia e ciência.	http://sedici.unlp.edu.ar/phpoai/oai2.php
The University of Pittsburgh. University Library System. Digital Research Library	A Biblioteca Digital da University of Pittsburgh's University Library System tem a missão de dar apoio na criação e disponibilização de coleções digitais acessíveis pela Web.	http://digital.library.pitt.edu/cgi-bin/b/broker20/broker20
Publicaties van de Universiteit van Amsterdam	Repositório de publicações eletrônicas da Universidade de Amsterdam, Holanda.	http://dare.uva.nl/cgi/arno/oai/uvapub
Oxford Eprints	Oxford E-prints é um repositório digital para consultas a artigos escritos por autores da Oxford University. A biblioteca conta com arquivos multidisciplinares, acessíveis a todos.	http://eprints.ouls.ox.ac.uk/perl/oai2
Nottingham eTheses	Arquivos eletrônicos de teses defendidas na University of Nottingham.	http://etheses.nottingham.ac.uk/perl/oai2
Middle East Technical University Library E-Thesis OAI Data Provider	Repositório que reúne teses vindas da Middle East Technical University Library.	http://etd.lib.metu.edu.tr/oai/oai2.php
Library of Congress Open Archive Initiative Repository	Coleção digital que reúne arquivos sobre a história Americana.	http://memory.loc.gov/cgi-bin/oai2
Les thèses en ligne de l'INP	Provedor de Dados de Teses eletrônicas na Internet.	http://ethesis.inp-toulouse.fr/perl/oai2
INRIA: ©HA	HAL-INRIA é uma plataforma que permite que pesquisadores pesquisem sobre documentos científicos diversos.	http://hal.inria.fr/oai/oai.php
Hiroshima University's Repository	Repositório eletrônico de documentos científicos disponíveis para consulta da Universidade de Hiroshima, Japão.	http://ir.lib.hiroshima-u.ac.jp/cgi-bin/oai/oai2.0

de todos os metadados de cada repositório. Este estudo de caso será detalhado nas próximas seções, dividas em módulos bem definidos para sua melhor compreensão.

6.1.1 Sistema de Administração

Ao entrar no módulo de Administração do Clio-i, é apresentada uma tela para o administrador logar no sistema. Digitando o login e a senha corretamente, é permitido o acesso à página principal deste módulo, como podemos ver na figura 6.1.

Figura 6.1 Telas Iniciais do Sistema de Administração.

Do lado direito da figura 6.1 encontramos seis ícones correspondentes a todas as funcionalidades disponíveis para um administrador do Clio-i. As seguintes opções são oferecidas:

- **Repositórios:** Realiza a inserção, exclusão, alteração e consulta de um repositório OAI.
- **Integrador de Dados:** Tem a função de coletar os metadados dos repositórios previamente cadastrados.
- **Mural de Recados:** Realiza a gerência dos recados postados por usuário no mural.
- **Administradores:** Inserção, exclusão, alteração e consulta de administradores do sistema Clio-i.
- **Configuração do Sistema:** Esta funcionalidade disponibiliza que o administrador realize pequenas modificações no sistema, como cor padrão das páginas, fonte e tamanho das letras, etc.

- **Fale Conosco:** Possui um formulário para postagem de dúvidas, sugestões e bugs do sistema para os desenvolvedores.

Como primeiro passo para a concretização do estudo de caso, precisamos realizar a inserção dos repositórios escolhidos na base de dados do Clio-i. A figura 6.2 apresenta a inserção do repositório *White Rose Consortium ePrints Repository*.

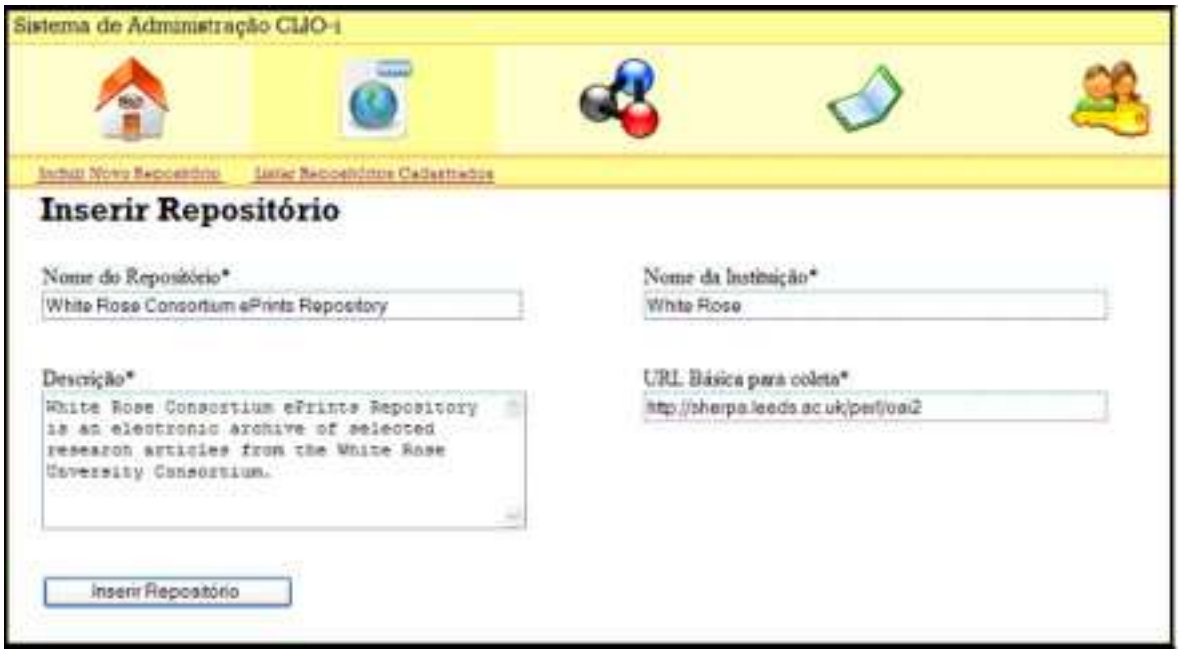

Figura 6.2 Inserção de um repositório no Clio-i.

Após a inserção de todos os Provedores de Dados OAI, damos início à coleta dos metadados de cada um deles, visto com detalhes na próxima seção.

6.1.2 Clio-i Service Provider

Ao clicar no ícone referente ao integrador de dados, é mostrada a lista de todos os repositórios cadastrados no sistema. Nesta mesma lista, é apresentado um campo para busca de um repositório específico, como mostra a figura 6.3.

Figura 6.3 Lista dos repositórios cadastrados no Clio-i.

Ainda de acordo com a figura 6.3, podemos perceber que abaixo da descrição do Provedor de Dados existem quatro ícones, que possuem as seguintes funções:

- **Mais detalhes:** Permite que o administrador realize a coleta dos metadados segundo alguns parâmetros (e.g. data e conjunto do metadado).
- **Coleta completa:** Realiza a coleta de todos os metadados do repositório cadastrado.
- **Cancelar coleta:** Permite que o administrador cancele a coleta de um repositório. Da próxima vez que essa coleta foi realizada, a mesma é iniciada a partir do ponto de onde é cancelada.
- **Deletar metadados:** Apaga todos os metadados coletados daquele repositório.

Na figura 6.4, é apresentada a coleta de todos os metadados do repositório White Rose Consortium ePrints Repository, acionado após clicarmos no ícone referente à coleta completa.

Depois de completada a coleta, uma mensagem é apresentada ao usuário, conforme mostra a figura 6.5.

Figura 6.4 Coletando os metadados de um Provedor de Dados.

Figura 6.5 Mensagem indicando sucesso na coleta dos dados.

Como já mencionado anteriormente, podemos realizar uma coleta detalhada de algum Provedor de Dados, através das datas do *datestamp* de um metadado coletado (detalhes sobre o *datestamp* no capítulo 3) e/ou de conjuntos ao qual um metadado pertence. Na figura 6.6 apresentamos a tela quando acionamos o ícone de mais detalhes, referente ao provedor *Université du Québec à Chicoutimi*.

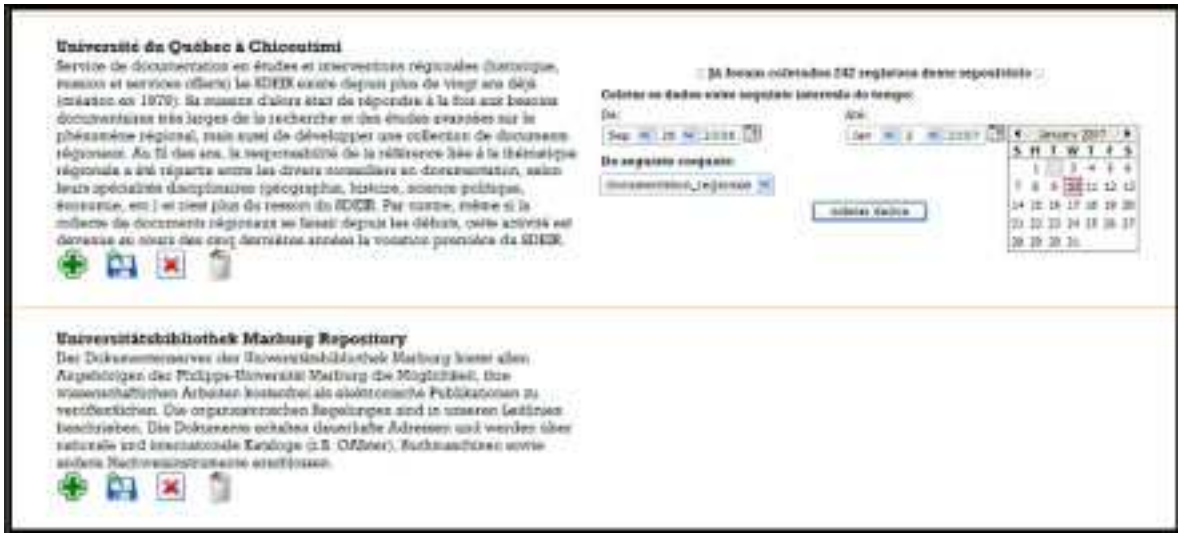

Figura 6.6 Realização da coleta a partir de refinamentos.

Do lado direito da descrição do Provedor de Dados em questão, encontra-se os detalhes de seus metadados. Primeiramente, percebemos que o sistema informa a quantidade de registros que já estão armazenados na base do Clio-i deste repositório específico (no caso, 242 registros coletados). Em seguida, o administrador pode selecionar o intervalo de tempo que deseja coletar os metadados, correspondente ao seu *datestamp*. As variáveis "De"e "Até"apresentadas neste tipo de coleta correspondem aos parâmetros *from* e *until*, respectivamente, do verbo *ListRecords* (detalhes no capítulo 3). Ao clicar na opção de mais detalhes, a data de início da coleta, correspondente ao *datestamp*, já aparece preenchida com o datestamp do último metadado coletado. Isso facilita a atualização do integrador de dados com metadados recentes de um repositório. Ainda sobre as datas, verificamos na figura o auxílio de um calendário (desenvolvido em DHTML) para escolher a data desejada, facilitando o administrador no momento de preencher esses campos.

Para finalizar os comentários sobre este tipo de coleta, o administrador pode selecionar o conjunto de metadados que queira coletar. Todos os conjuntos do repositório são solicitados através do verbo *ListSets* (detalhes no capítulo 3) cada vez que a opção de mais detalhes para coleta é requisitada. Na coleta dos dados, esse conjunto corresponde ao parâmetro set do verbo *ListRecords* (detalhes também no capítulo 3).

Após a coleta de todos os metadados dos dezenove repositórios selecionados, reunimos 138.510 registros científicos reunidos na base de dados do Clio-i. A tabela 6.2 apresenta detalhes da coleta de cada Provedor de Dados, informando a quantidade de registros coletados e o

tempo necessário para a tarefa.

Tabela 6.2 Relatório de coleta dos metadados

Provedor de Dados OAI	**Quantidade de Registros Coletados**	**Tempo gasto(em minutos)**
White Rose Consortium ePrints Repository	1.737	10
ViFaPhys.de	426	6
Université du Québec à Chicoutimi - Documentation régionale	242	5
Universitätsbibliothek Marburg	1.228	19
Universität Karlsruhe: EVA - Elektronisches Volltextarchiv	535	5
University of Wuerzburg, GERMANY, OPUS	1.777	30
University of Wisconsin Digital Collections	33.500	85
University of Pittsburgh Electronic Thesis and Dissertation Archive	1.359	10
University of Joensuu: electronic publications	354	2
Universidad Nacional de La Plata	1.050	8
The University of Pittsburgh. University Library System. Digital Research Library	15.639	38
Publicaties van de Universiteit van Amsterdam	63.200	368
Oxford Eprints	553	9
Nottingham eTheses	106	1
Middle East Technical University Library E-Thesis OAI Data Provider	449	4
Library of Congress Open Archive Initiative Repository	600	7
Les thèses en ligne de l'INP	211	2
INRIA: ©HA	7.926	62
Hiroshima University's Repository	7.618	76

Como podemos perceber, houve uma variação muito grande entre número de registros coletados e o tempo gasto para coletá-los. Repositórios como *University of Wisconsin Digital Collections*, que possuem mais de 33 mil documentos à disposição para a coleta, tiveram uma taxa de transferência de aproximadamente 394 registros por minuto. Por outro lado, o Provedor de Dados *ViFaPhys.de*, onde foram coletados apenas 426 registros, obteve uma taxa de transferência bem menor, em torno de 71 registros por minuto. Na média, para coletar todos os 138.510 registros, o Clio-i obteve uma taxa de aproximadamente 185 registros de metadados por minuto.

Evidentemente, o tempo para coleta dos registros depende bastante de onde os sistemas (tanto o Provedor de Serviços, quando o Provedor de Dados) estão alocados. O Clio-i Service Provider estava funcionando em um computador pessoal (1.46 GHz, 500 MB de RAM),

conectado na rede a uma velocidade de 48,00 Mbps para realizar a coleta destes registros.

6.1.3 Módulo de Recuperação de Informação

A página inicial do Clio-i deste estudo de caso pode ser vista na figura 6.7.

Figura 6.7 Página principal do Integrador de Repositórios Científicos.

Na área superior da figura 6.7, encontram-se ícones que mostram as opções que os usuários podem utilizar (e.g. busca avançada, mural de recados). Do lado esquerdo, encontramos a lista dos 19 repositórios e o número de registros coletados de cada um destacado entre parênteses. Na área central, é apresentado o módulo de Recuperação da Informação, com um campo para o usuário realizar a pesquisa nos 138.510 registros inseridos na base.

Um exemplo deste módulo pode ser encontrado na figura 6.8, que apresenta o resultado de uma consulta com o termo "Brazil". Neste resultado, são mostrados todos os metadados coletados do registro e o termo pesquisado aparece destacado.

Figura 6.8 Resultado da pesquisa no Clio-i.

Vale ressaltar a qualidade da consulta no que diz respeito à ordenação dos registros retornados. Como podemos perceber, ainda na figura 6.8 são apresentados os primeiros dos quarenta e seis registros encontrados com a palavra "Brazil". Esses registros possuem uma quantidade considerável da palavra consultada, mostrando a relevância de cada um sobre o tema específico. O termo requisitado, entretanto, não aparecia mais do que uma vez entre os metadados dos últimos registros da consulta.

Ao clicarmos no nome de algum repositório, localizado na parte esquerda da página, é possível realizar a consulta dentro dos metadados coletados do repositório selecionado. Na figura 6.9, encontramos um exemplo de uma consulta, com a mesma palavra-chave "Brazil", realizada no Provedor de Dados *University of Wisconsin Digital Collection*.

Figura 6.9 Consulta em um repositório específico.

Ressaltamos aqui que o Visualizador de Documentos não é utilizado neste estudo de caso, visto que os Provedores de Dados selecionados não disponibilizam o recurso eletrônico de forma a atender a esse requisito. No próximo estudo de caso, utilizaremos como Provedores de Dados alguns Clio-i Data Providers e, com isso, demonstraremos a Visualização de Documentos em outras bases.

6.1.4 Clio-i Data Provider

Relatamos em capítulos anteriores que seria de grande importância que um Provedor de Serviços que mantivesse uma base única de dados coletados de diversas fontes também exportasse esses dados, agindo também como um Provedor de Dados.

Desta maneira, o Clio-i atende a esse requisito adicionando ao sistema um módulo do Clio-i Data Provider. A figura 6.10 apresenta a resposta em XML da seguinte requisição:

http://www.liber.ufpe.br/clioi/modules/pd/index.php?
verb=ListRecords&metadataPrefix=oai_dc

Figura 6.10 Resposta em XML da requisição HTTP.

A base de dados coleta nesse estudo de caso foi registrada oficialmente como Provedor de Dados no *Open Archives Initiative*. É o terceiro Provedor de Dados brasileiro registrado, mas o único até o momento que estava em pleno funcionamento. Na figura 6.11, é mostrado uma parte da lista dos Provedores de Dados oficiais OAI, com o Clio-i (deste estudo de caso) em destaque.

Figura 6.11 Clio-i Data Provider oficialmente registrado na OAI.

Para concretizarmos esse registro, realizamos uma requisição ao site do OAI, enviando a URL básica para coleta do sistema. A partir daí, automaticamente o Provedor de Dados deste estudo de caso foi testado exaustivamente e, como não houve nenhuma inconformidade com o protocolo, o repositório foi registrado na lista do *Open Archives*.

6.2 Estudo de Caso 2: Integrador de Repositórios Multimídia

No primeiro estudo de caso, demonstramos a potencialidade do Clio-i Service Provider para coletar bases de dados de repositórios que seguem o protocolo padrão OAI-PMH. Neste segundo estudo de caso, pretendemos avaliar mais especificamente o módulo Clio-i Data Provider com a extensão do protocolo OAI-PMH. Apresentamos aqui três projetos de Bibliotecas Digitais que usaram o Clio-i, e utilizamos o Clio-i Data Provider para expor os metadados desses projetos. Os três provedores de dados foram utilizados nos seguintes projetos:

- *Acervo Digital FUNDAJ:* Já discutido anteriormente no capítulo 4, o acervo digital da Fundação Joaquim Nabuco utiliza os serviços de Recuperação da Informação e Visualização de Documentos do Clio-i desde o início do ano de 2006. Adicionamos ao sistema o módulo para expor os seus metadados e os seus recursos, através do Clio-i Data Provider.

- *Holandeses na Bahia:* Base composta essencialmente por documentos do tipo texto, sobre a conquista holandesa de Salvador no Brasil em 1624, e a reconquista da cidade por uma armada luso-espanhola em 1625.

- *Escrito nas Estrelas:* Criado no Laboratório Liber-UFPE, o projeto Escrito nas Estrelas realiza uma exposição sobre o Brasil na época do desenvolvimento da aviação, dos transportes aéreos, dos correios, e o impacto da nova tecnologia na forma de ver e pensar do homem do início do século XX. Para o presente trabalho, foram selecionados alguns vídeos do projeto que serão utilizados em um Clio-i Data Provider.

Os metadados e os registros dessas três bases foram integrados por um Clio-i Service Provider, na qual chamamos o projeto de Integrador de Repositórios Multimídia. Se por um lado este projeto agrega uma quantidade de documentos muito pequena em relação ao primeiro estudo de caso, por outro demonstramos a facilidade em se trabalhar com os quatro tipos de mídias suportadas pelo Clio-i, além de apresentarmos a extensão do protocolo na prática. As próximas três subseções apresentarão os Provedores de Dados utilizados.

6.2.1 Acervo Digital FUNDAJ

O Acervo Digital da Fundação Joaquim Nabuco disponibiliza documentos digitalizados mantidos nos centros de documentação da instituição. Atualmente, o acervo conta com 1.010 documentos, todos do tipo imagem, que estão categorizados em diversas coleções (e.g.: Cordéis, Rótulos de Cigarro, Cartões-Postais, etc.). Através dos serviços oferecidos pelo Clio-i esses documentos são disponibilizados aos usuários e por meio do Clio-i Data Provider os seus metadados, assim como os recursos, são exportados para futura coleta.

Um dos serviços que é disponibilizado é o sistema de busca avançado, conforme mostra a figura 6.12. No caso, o usuário pode especializar a sua consulta através de operados booleanos (e.g.: OR, AND e NOT), especificar o idioma que deseja procurar e a coleção que o documento pertence (na figura, especificado como projeto).

Ainda temos o sistema de administração, responsável, dentre outras funções (todas mencionadas no capítulo 4), de inserir documentos na base. A inserção de qualquer documento no Clio-i é realizada através de duas etapas. Na primeira delas, o administrador do sistema insere todos os metadados referentes ao recurso. Após isso, é apresentada a segunda etapa, referente à inserção do recurso em si. Na figura 6.13, identificamos a primeira etapa desta inserção. Nela, o administrador deve preencher a coleção a qual o documento pertence, qual o seu tipo de mídia e os seus metadados.

Figura 6.12 Busca Avançada.

Figura 6.13 Primeiro passo para a inserção dos documentos.

Na segunda etapa da funcionalidade, inserimos os documentos do tipo imagem na base, apresentado na figura 6.14.

Figura 6.14 Inserindo as imagens do documento.

Por fim, este Provedor de Dados, assim como os outros dois que serão apresentados mais a frente, exporta seus dados em formato de XML, de acordo com a extensão do protocolo OAI-PMH. Disponibiliza assim, o recurso eletrônico de cada registro, além de dar suporte ao parâmetro *query*, e ao verbo *GetRecord*, explicados com detalhes no capítulo 4. Se realizarmos a seguinte requisição a este repositório:

http://digitalizacao.fundaj.gov.br/fundaj2/modules/pd/
index.php?verb=ListRecords&metadataPrefix=oai_dc

Iremos obter uma reposta em XML de acordo com a figura 6.15, mostrando a tag *dc:cliodocumento*, informando o caminho físico de cada imagem do recurso.

Figura 6.15 Resposta XML no Acervo Digital FUNDAJ.

Assim como realizado no primeiro estudo de caso, decidimos tornar o Provedor de Dados da Fundação Joaquim Nabuco oficialmente registrado no *Open Archives Initiative*. Após a avaliação da OAI, o repositório foi aceito e também se encontra na lista oficial da entidade.

6.2.2 Holandeses na Bahia

O segundo projeto, usado no estudo de caso, dispõe de documentos essencialmente do tipo texto, sobre o período em que a Bahia esteve conquistada pelos holandeses. Os documentos contidos nessa base estão divididos basicamente em três coleções:

- **Cadena Vilhasanti:** Reúne documentos escritos por Pedro Cadena de Vilhasantie, provedor-mor da Bahia na época da chegada dos holandeses em 1624.

- **Jornada dos Vassalos:** Reúne relatos do pesquisador da época, Mattheus Pinheiro, sobre Dom Fradique, comandante da armada espanhola na retomada da Bahia.

- **Tomada da Bahia:** Cartas de Henrique Moniz Telles a respeito da retomada da Bahia, em 1638.

A inserção dos documentos é bem parecida com a mostrada no Acervo Digital da FUNDAJ. Entretanto, como se trata de documentos do tipo texto, a única diferença é identificada na segunda etapa da inserção, como mostra a figura 6.16.

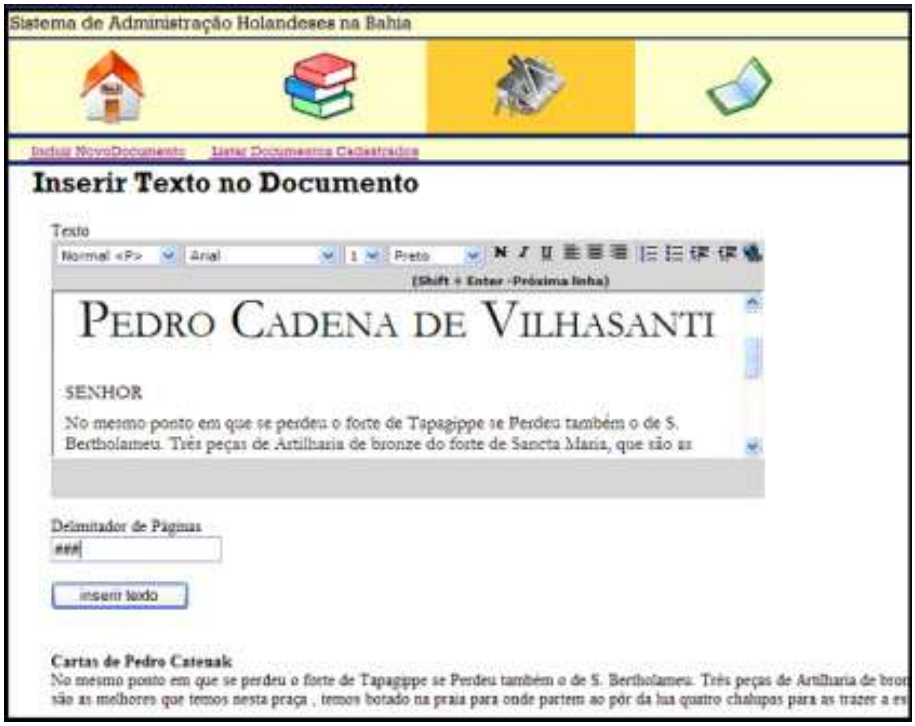

Figura 6.16 Inserindo textos no documento.

Como podemos perceber, há um espaço para se colocar o texto do documento, na qual é mantida a sua formatação original (e.g.: tamanho, cor e tipo da fonte, espaçamento, alinhamento, etc.). Há ainda uma outra opção, a de delimitarmos as páginas de um documento, no caso da figura 6.16, o símbolo "###". Este símbolo irá indicar ao sistema que uma nova página se inicia.

O texto inserido pode então ser visto pelo usuário através do Visualizador de Documentos, conforme mostra a figura 6.17.

Figura 6.17 Visualizador de Documentos com arquivo do tipo texto.

6.2.3 Escrito nas Estrelas

O último repositório criado para este estudo de caso trata de documentos no formato vídeo, vindos do projeto Escrito nas Estrelas. A base conta com algumas dezenas de vídeos que testemunham os principais momentos da passagem do piloto José Manoel Sarmento Beires no Brasil.

Este repositório, assim como os outros dois mostrados anteriormente, utiliza o Clio-i para operacionalizar as suas informações. Assim, as suas funcionalidades são idênticas às mostradas anteriormente. Vale ressaltar, entretanto, o suporte às mídias vídeo e áudio, que utiliza um *plug-in* adequado (e.g., Windows Media Player, QuickTime) executar estes tipos de arquivos. Na figura 6.18 é apresentado o Visualizador de Documentos mostrando um vídeo deste projeto.

Figura 6.18 Vídeo no Clio-i.

6.2.4 Integrador de Repositórios Multimídia

O Integrador de Repositórios Multimídia deste segundo estudo de caso realiza a coleta dos três repositórios mostrados anteriormente, integrando não só os seus metadados, como também os recursos eletrônicos disponibilizados em cada base. Evidentemente, por se tratar do mesmo sistema, tudo o que foi mostrado no primeiro estudo de caso, também foi aplicado neste segundo (e.g.: busca nos documentos, coleta de informações dos repositórios, etc.). Assim, na figura 6.19, encontramos o resultado de uma consulta realizada com o termo "bahia"sobre os documentos coletados das três bases criadas.

De acordo com a figura 6.19, podemos perceber que a única diferença entre este resultado e mostrado na figura 6.8, é a presença de ícones indicando que o usuário pode visualizar o recurso eletrônico por completo, através do Visualizador de Documentos. Ainda sobre a figura 6.19, o primeiro registro que é apresentado é uma imagem coletada do Acervo Digital da FUNDAJ. Já o segundo registro corresponde a 10 páginas de texto oriundas do repositório Holandeses na Bahia.

Figura 6.19 Página inicial do Integrador de Repositórios Multimídia.

Outra especificidade encontrada neste estudo de caso é a utilização da extensão do protocolo OAI-PMH. Como exemplo, demonstramos a figura 6.20, que apresenta a tela principal para a coleta de informações.

Figura 6.20 Coleta de informações com extensão do protocolo OAI-PMH.

Nesta figura, percebemos alguns detalhes não presentes na coleta do primeiro estudo de caso. O primeiro deles é o uso do verbo *GetSize*, explicado com detalhes no capítulo 4, que aparece na figura informando a quantidade de registros em um dado repositório. Após isso, percebemos um campo identificado por termo. Este campo é necessário para o administrador coletar informações que contenham determinadas palavras-chave em seus metadados, utilizado no parâmetro *query* do verbo *ListRecords*.

Além dessas modificações visíveis na figura 6.20, também é realizado a coleta do objeto eletrônico, identificado pela tag *dc:cliodocument*, encontrada nos metadados de um recurso. Vale ressaltar que os recursos do tipo texto são inseridos na base de dados por completo, enquanto os outros tipos (áudio, vídeo e imagem), apenas o caminho onde o mesmo se encontra é armazenado para serem usados posteriormente no Visualizador de Documentos.

6.3 Considerações Finais

Os estudos de caso apresentados no capítulo demonstram todas as funcionalidades que acreditamos serem importantes para sistemas de Bibliotecas Digitais. O sistema de Recuperação de Informação apresentou uma performance bastante aceitável, tanto em tempo de resposta quanto em qualidade na recuperação da informação. Características essas mostradas em uma base com centenas de milhares de registros. Além disso, foi apresentado o módulo Visualizador de Documentos, que suporta a operacionalização de quatro tipos de mídias diferentes.

Apresentamos ainda a capacidade do Clio-i de integrar Bibliotecas Digitais, conforme visto nos dois estudos de caso. Para isso, realizamos a coleta eficiente de milhares de registros de Provedores de Dados OAI com o protocolo padrão (primeiro estudo de caso), e ainda a coleta de informações de Clio-i Data Providers conforme a extensão do protocolo (segundo estudo de caso)

Por se tratar de um protótipo, ainda há muito a se fazer, tanto em adaptações do sistema a novas necessidades, quanto aperfeiçoamento das que já existem. Isso tudo será discutido no capítulo seguinte, onde apontaremos os trabalhos futuros, realizando a conclusão de nossas pesquisas.

CAPÍTULO 7

Conclusões

O virtual, rigorosamente definido, tem somente uma pequena afinidade com o falso, o ilusório ou o imaginário. Trata-se, ao contrário, de um modo de ser fecundo e poderoso, que põe em jogo processos de criação, abre futuros, perfura poços de sentido.

—PIERRE LÉVY

Neste trabalho, investigamos diferentes serviços de Bibliotecas Digitais, visando identificar limitações dos projetos da área. A maioria dos sistemas não possuía um engenho de busca qualificado e só apresentava os metadados descritivos do recurso, não o seu conteúdo completo. Além disso, para uma eficiente disseminação da informação, é necessário aplicar às Bibliotecas Digitais, padrões que promovem a integração entre bases na Internet. Com a visão voltada para esta carência e a sua crescente demanda, foi projetado e implementado o sistema Clio-i.

Destaca-se no presente trabalho, a facilidade de interoperabilidade do sistema com outros repositórios que estejam de acordo com o protocolo adotado, o OAI-PMH. Avaliamos a capacidade do sistema de integrar dados, em um estudo de caso onde milhares de registros de metadados foram coletados de diferentes repositórios científicos e armazenados em uma base de dados centralizada. Visando inserir funcionalidades ao protocolo padrão, desenvolvemos nesse trabalho uma extensão do OAI-PMH, e realizamos um estudo de caso integrando as Bibliotecas Digitais de três diferentes projetos.

A seguir apresentamos um resumo das principais contribuições do nosso trabalho (seção 7.1) e alguns trabalhos futuros (seção 7.2). As considerações finais serão feitas na seção 7.3.

7.1 Resumo das Contribuições

A seguir apresentamos um resumo das principais contribuições do nosso trabalho durante o mestrado.

- **Critérios de avaliação:** A partir de pesquisas realizadas em diferentes Bibliotecas Digitais, defimos critérios para a avaliação de uma Biblioteca Digital, impulsionado pela demanda atual e pela qualidade que sistemas na área devem oferecer aos seus usuários.

- **Detalhamento do *Open Archives Initiative:*** Durante a nossa dissertação, reservamos o capítulo 3 para realizarmos um comentário detalhado sobre o *Open Archives Initiative*. Tecemos considerações sobre o seu protocolo, o OAI-PMH e definimos alguns critérios de avaliação para sistemas que utilizam o OAI para interoperabilidade de seus dados.

- **Extensão do protocolo OAI-PMH:** Desde o seu surgimento, em meados de 2001, o protocolo OAI-PMH foi bastante aceito pela comunidade especializada em integração de dados na Internet. Essa grande aceitação se deu pelo fato do protocolo trabalhar com tecnologias já bastante difundidas, como o HTTP, o XML e o padrão de metadados Dublin Core. Entretanto, durante nossa pesquisa, algumas necessidades para a integração entre repositório foram percebidas que o protocolo do OAI não atendia. Sendo assim, realizamos uma extensão do protocolo OAI-PMH e a aplicamos no sistema Clio-i.

- **Proposta e implementação do sistema Clio-i:** A fim de suprir as necessidades encontradas e a grande demanda percebida, realizamos a proposta do sistema Clio-i. Para isso, desenhamos a sua arquitetura modularizada baseada em componentes, responsáveis pela recuperação e visualização da informação, e pela interoperabilidade entre bases na Internet. O passo seguinte foi a sua implementação, validada com testes funcionais e de usabilidade.

- **Montagem de Estudos de Caso:** A fim de demonstrar toda a potencialidade do Clio-i, dois estudos de caso foram construídos. O primeiro realizou a coleta de centenas de milhares de registros de dezenove diferentes repositórios, que foram agrupados em uma base de dados única. O segundo estudo de caso reuniu documentos em diversos formatos (e.g.: vídeo, imagem e texto) e apresentou na prática a extensão do protocolo OAI-PMH adotada.

7.2 Trabalhos Futuros

O sistema proposto tem a facilidade de trabalhar com diversas mídias, especificamente áudio, texto, vídeo e imagem. Um trabalho a médio prazo que pode ser estabelecido é permitir que o Clio-i operacionalize com outras mídias, ou com informações que possuam mais de uma mídia (e.g. notícias que possuam texto e vídeo, imagem com áudio associado, etc.).

O sistema, antes mesmo de finalizado, já despertou interesse em algumas instituições, tanto nacionais quanto internacionais. Desta maneira, a curto prazo, iremos montar uma equipe para realizarmos mais testes no sistema e realizarmos os ajustes necessários para a sua utilização.

A médio prazo, estaremos pesquisando mais necessidades sobre a interoperabilidade entre bases e, por conseguinte, expandindo o OAI-PMH na medida que for possível. Como trabalho futuro também destacamos a necessidade da utilização do OAI-PMH, com outros protocolos de interoperabilidade de dados, como o Z30.50 [Lyn97], aumentando assim o poder de comunicação entre bases diversas na Internet.

Para finalizar, pretendemos estudar a fundo diversos protocolos para interoperabilidade entre Bibliotecas Digitais, detalhando seus benefícios e suas carências. Com este estudo, podemos criar um protocolo próprio para comunicação entre bases na Internet, direcionando-o para especificações de nossa escolha.

7.3 Considerações Finais

O estudo sobre Bibliotecas Digitais, assim como a sua interoperabilidade com outras bases, é um tema que vem crescendo de importância nos últimos anos, devido à grande proliferação da informação na Web. No presente trabalho, procuramos dar a nossa contribuição na área, apresentando um estudo detalhado sobre o tema e criando um sistema para atender a essa crescente demanda.

Nossa intenção é que o Clio-i atenda a uma grande quantidade de instituições, realizando críticas ao sistema que nos permitam torna-lo cada vez mais eficiente na área. Esperamos ainda que a dissertação enriqueça a quantidade de informação na área de Bibliotecas Digitais e integração de dados, servindo de consulta para pesquisadores interessados no assunto.

APÊNDICE A

Relatório de Avaliação dos Testes Funcionais

Este apêndice tem o objetivo sumarizar os testes funcionais realizados no Clio-i, apresentado os resultados e modificações realizadas no sistema para suprir qualquer inconformidade apresentada. Na tabela A1, apresentamos o sumário dos resultados de todos os casos de testes realizados.

Tabela A.1 Sumário dos resultados dos casos de testes

Caso de Testes	**Status**	**Resultado**	**Registro de Incidentes**
[CT01] Pesquisar Documento	Realizado	Sucesso	
[CT02] Visualizar Documento	Realizado	Sucesso	
[CT03] Cadastrar Coleção	Realizado	Sucesso	
[CT04] Cadastrar Documento	Realizado	Sucesso	
[CT05] Cadastrar Repositório	Realizado	Sucesso	
[CT06] Coletar Metadados	Realizado	Insucesso	RI06

Apenas um incidente foi registrado em todos os casos de testes, durante a coleta dos Provedores de Dados OAI. No quadro A.2, encontramos o resumo do incidente e quais medidas foram tomadas para saná-lo.

Tabela A.2 Registro do Incidente RI06

[CT01] Pesquisar Documento	
Caso de Teste	[CT06] Coletar Metadados
Descrição da Falha	O Integrador de Dados do Clio-i não realizava a coleta dos Provedores de Dados que possuíssem a data no formato YYYY-MM-DDThh:mm:ssZ.
Correções Necessárias	Incluir suporte a datas dos Provedores de Dados no formato YYYY-MM-DDThh:mm:ssZ.
Resultado	Falha sanada com sucesso.

APÊNDICE B

Plano de Teste de Usabilidade

B.1 Propósito do Teste

O propósito deste teste é verificar a performance alcançada pelos participantes e o entendimento das funções do sistema utilizando o protótipo, com a finalidade de realizar alterações necessárias para atender à demanda de usuários do Clio-i. Será medido o tempo gasto para a realização das tarefas e serão identificados erros e dificuldades envolvendo a utilização do protótipo em tarefas rotineiras.

B.2 Declaração dos Problemas

Pretendemos com este teste de Usabilidade responder a duas questões básicas: (1) os termos utilizados nas interfaces são intuitivos? (2) o desempenho alcançado pelos usuários é o ideal?

B.3 Perfil do Usuário

Serão utilizados quatro participantes, dois por dia. Os participantes devem ter de 20 a 40 anos de idade, nível médio (completo ou não) ou superior (completo ou não), mais de um ano de conhecimentos básicos de informática e de utilização de aplicativos básicos e Internet.

B.4 Metodologia

O teste será realizado com a finalidade de garantir a usabilidade do produto e será composto das seguintes partes:

- Cada participante será devidamente cumprimentado pelo avaliador, será orientado a se sentar e tentar se sentir confortável e relaxado. O participante será orientado a preencher um pequeno questionário para identificação de seu perfil.
- O participante receberá um script introdutório de orientação do teste, explicando o propósito e objetivos do teste e o que é esperado dos participantes. Deve ser reforçado que o produto é o centro da avaliação e não o participante e que as tarefas devem ser executadas de forma bastante confortável. Deve-se informar ao participante que ele será observado em todas as suas tarefas pelo avaliador.

- Depois de passadas as orientações, será permitido que o participante utilize o sistema livremente por cinco minutos. Logo depois, será entregue a lista de tarefas (Lista de Tarefas, Apêndice C). O avaliador irá requisitar que o participante verbalize suas dúvidas, pois isto ajudará ao avaliador anotar a ocorrência e a razão de problemas. Durante o teste, os acontecimentos observados pelo avaliador serão registrados em formulário próprio.
- Depois de completadas todas as tarefas, o participante preencherá um questionário de avaliação do sistema pelo participante cuja finalidade é coletar informações preferenciais do participante (Questionário de Avaliação do Sistema pelo Participante, Apêndice D).
- Depois, o participante será questionado pelo avaliador em uma sessão de questionamento do participante. Serão discutidas percepções subjetivas de usabilidade do participante acerca do sistema, realizados comentários globais sobre a performance do participante e problemas encontrados. O participante poderá comentar sobre o teste abertamente, permitindo uma coleta de informações complementares.
- A sessão de questionamentos se encerra e a sua colaboração é agradecida.

B.5 Papel do Avaliador

O avaliador se sentará ao lado do participante durante a realização do teste e registrará o tempo gasto nas tarefas, erros e observações através de um formulário.

O avaliador não poderá ajudar o participante na realização das tarefas. Ele somente poderá orientar se surgir uma questão acerca do procedimento de teste. Um ajudante do avaliador irá cronometrar e registrar o tempo gasto na realização das tarefas.

B.6 Medidas de avaliação

As seguintes medidas de avaliação serão coletadas e calculadas:

- Tempo gasto para completar cada tarefa por participante.
- Número de erros cometidos na realização de cada tarefa por participante.
- Tempo médio gasto na execução de cada tarefa.
- Média de erros por tarefa.
- Dados qualitativos sobre a utilização do protótipo do sistema Clio-i.

B.7 Conteúdo do Relatório e Apresentação

O relatório irá conter o plano de testes, resultados, discussões e recomendações. Os resultados finais serão compostos de itens e recomendações que serão apresentados aproximadamente alguns dias depois do teste. Incluirá revisões preliminares a fim de completar a análise proposta.

APÊNDICE C

Relatório de Avaliação dos Testes Funcionais

Agora, você dará início aos testes.

Abaixo, nós temos vinte e duas tarefas, divididas em três grupos principais, que devem ser executadas por você utilizando o produto.

As tarefas devem ser executadas na ordem em que se encontram. Você deve ler em voz alta cada tarefa antes de executá-la.

Lembre-se:

- Utilize o sistema livremente por 5 minutos, em cada um dos três grupos.
- Verbalize suas dúvidas, pois isto ajudará ao avaliador anotar a ocorrência e a razão de problemas.
- É o produto que está sendo avaliado e não você.

C.1 Grupo A - Sistema para uso dos Usuários Comuns

Neste grupo você irá testar os itens que estão disponíveis para qualquer usuário que acessa o Clio-i pela Internet. Especificamente falando, você estará acessando a documentos eletrônicos contidos no acervo digital da Fundação Joaquim Nabuco.

- **Tarefa 1:** Você tem o interesse sobre a escravidão no Brasil. Realize uma pesquisa para encontrar documentos sobre este assunto.
- **Tarefa 2:** Tente visualizar o documento em questão.
- **Tarefa 3:** Percorra as páginas do documento aberto.
- **Tarefa 4:** Com o documento visualizado, aumente-o, utilizando as opções disponíveis.
- **Tarefa 5:** Com o mesmo documento, aplique um efeito de negativação.
- **Tarefa 6:** Insira alguma nota sobre o documento.
- **Tarefa 7:** Realize o Download do documento para sua máquina.
- **Tarefa 8:** Saia da Visualização do Documento.
- **Tarefa 9:** Vá para a opção Pesquisa Avançada.
- **Tarefa 10:** Realize uma pesquisa qualquer apenas na coleção Cordéis.

C.2 Grupo B - Sistema Administrativo 1: Inserção de Documentos na base

Neste grupo você irá testar os itens que estão disponíveis apenas para administradores do Clio-i. Novamente, você estará acessando ao sistema administrativo do acervo digital da Fundação Joaquim Nabuco.

- **Tarefa 11:** Logue-se no sistema administrativo. O seu login será "teste"e a sua senha de acesso será "usabilidade".

- **Tarefa 12:** Insira algum documento do tipo imagem na base de dados. O valor dos campos referentes ao documento pode ser de sua escolha. As imagens correspondentes ao documento encontra-se no caminho "C:
  Testes
  Documento
  ".

- **Tarefa 13:** Verifique se o documento foi inserido com sucesso, realizando uma pesquisa do mesmo.

- **Tarefa 14:** Altere o documento inserido, apagando a sua última página.

- **Tarefa 15:** Exclua o documento inserido.

C.3 Grupo C - Sistema Administrativo 2: Coleta de documentos de outros repositórios

Neste grupo, novamente você irá testar os itens que estão disponíveis apenas para administradores do Clio-i. Desta vez, iremos realizar a coleta de documentos de instituições que estão localizadas em algum lugar da Internet. Você estará utilizando o sistema dentro do projeto chamado "Integrador de repositórios científicos", que reúne repositórios científicos em um único lugar para acesso a usuários.

- **Tarefa 16:** Logue-se no sistema administrativo. O seu login será "teste"e a sua senha de acesso será "usabilidade".

- **Tarefa 17:** Insira um repositório na base de dados. O valor dos campos referentes ao repositório serão os seguintes:
  - Nome do Repositório: ViFaPhys.de
  - Nome da Instituição: ViFaPhys
  - URL Básica para coleta: http://vifaphys.tib.uni-hannover.de/oai/oai2.php
  - Descrição: *The aim of the Physics Virtual Library ViFaPhys is to provide a central access point to information and services relevant to physics. The Subject Guide is a clearly organised, up-to-date compilation of selected and quality-controlled information*

sources. Short descriptions characterise contents and services of each resource. Included resources are accessible mainly via the Internet.

- **Tarefa 18:** Realize a coleta de todos os Metadados do Repositório.
- **Tarefa 19:** Após a coleta, apague todos os metadados que foram coletados..
- **Tarefa 20:** Colete novamente os metadados do Repositório, agora especificando que os metadados coletados devem compreender entre as datas 25/10/2000 até 30/01/2005.
- **Tarefa 21:** Novamente, apague todos os documentos coletados.
- **Tarefa 22:** Apague o repositório inserido anteriormente.

APÊNDICE D

Questionário de Avaliação do Sistema pelo Participante

O objetivo deste questionário é colher informações sobre a opinião do participante do teste de usabilidade que foi realizado utilizando o protótipo do Clio-i. As informações fornecidas são vitais para o aprimoramento do sistema.

Nas questões, favor circular o número correspondente ao grau de concordância.

Por favor, leia com atenção as questões a seguir e em caso de dúvida, solicite esclarecimento com o avaliador.

Tabela D.1 Questionário de avaliação

a.	Facilidade de utilização	0 1 2 3 4 5
b.	Organização das informações	0 1 2 3 4 5
c.	Layout das telas	0 1 2 3 4 5
d.	Nomenclatura utilizada nas telas (nome de comandos, títulos, campos, etc.)	0 1 2 3 4 5
e.	Mensagens do sistema	0 1 2 3 4 5
f.	Assimilação das informações	0 1 2 3 4 5
g.	Satisfação na realização dos testes	0 1 2 3 4 5

Referências Bibliográficas

[ABS00] Serge Abiteboul, Peter Buneman, and Dan Suciu. *Data on the Web: from relations to semistructured data and XML*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2000.

[Arm00] W.Y Arms. Digital libraries. *MIT Press*, 2000.

[Atk98] Dan Atkins. Vision for digital libraries. *An International research agenda for digital*, pages 11–14, October 1998.

[BDF+01] Peter Buneman, Susan B. Davidson, Wenfei Fan, Carmem S. Hara, and Wang Chiew Tan. Keys for XML. In *World Wide Web*, pages 201–210, 2001.

[Braa] Clube OAI Brasil. Bibliotecas temáticas de acesso aberto. Disponível em: http://clube-oai.incubadora.fapesp.br/portal/prot-oai/tematicas. Acesso em: 29 maio 2006.

[Brab] Clube OAI Brasil. Comunidade clube oai brasil. Disponível em: http://clube-oai.incubadora.fapesp.br/portal. Acesso em: 21 nov. 2006.

[Brac] Comunidade OAI Brasil. E-prints oai brasil. Disponível em: http://clube-oai.incubadora.fapesp.br/portal/softwares/eprints. Acesso em 12 jun. 2006.

[BTW03] David Bainbridge, John Thompson, and Ian H. Witten. Assembling and enriching digital library collections. In *JCDL '03: Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries*, pages 323–334, Washington, DC, USA, 2003. IEEE Computer Society.

[BYRN99] Ricardo A. Baeza-Yates and Berthier A. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press / Addison-Wesley, 1999.

[Car05] Marcos Cardoso. Bibliotecas digitais para documentos históricos. estudo de caso: Memórias do golpe: O brasil de 64 a 85, 2005. Trabalho de Graduação em Ciência da Computação. Centro de Informática. Universidade Federal de Pernambuco.

[CBG+05] Marcos Cardoso, Flávia Barros, Marcos Galindo, Ricardo Prudêncio, and Mercedes Otero. Clio: Um sistema de gerenciamento de bibliotecas digitais para documentos históricos. *3o. Simpósio Internacional de Bibliotecas Digitais*, November 2005.

[CDD+96] David M. Choy, Richard Dievendorff, Cynthia Dwork, Jeffrey B. Lotspiech, Robert J. T. Morris, Norman J. Pass, Laura C. Anderson, Alan E. Bell, Stephen K. Boyer, Thomas D. Griffin, Bruce A. Hoenig, James M. McCrossin, Alex M. Miller, Florian Pestoni, and Deidra S. Picciano. The almaden distributed digital library system. In *ADL '95: Selected Papers from the Digital Libraries, Research and Technology Advances*, pages 203–220, London, UK, 1996. Springer-Verlag.

[CEG+99] Yuan Chen, Jan Edler, Andrew Goldberg, Allan Gottlieb, Sumeet Sobti, and Peter Yianilos. A prototype implementation of archival intermemory. In *Proceedings of the Fourth ACM International Conference on Digital Libraries*, 1999.

[Cen] PubMed Central. Pubmed central overview. Disponível em: http://www.pubmedcentral.nih.gov/about/intro.html. Acesso em 13 jun. 2006.

[Cora] Dublin Core. Dublin core elements. Disponível em: http://www.dublincore.org/documents/dces/. Acesso em: 14 set. 2006.

[Corb] Dublin Core. Dublin core metadata initiative. Disponível em: http://www.dublincore.org. Acesso em: 14 out. 2006.

[Corc] Dublin Core. *User Guide Dublin Core*. Disponível em: http://www.dublincore.org/documents/usageguide/. Acesso em: 14 set. 2006.

[CW03] Gregory Crane and Clifford Wulfman. Towards a cultural heritage digital library. In *JCDL '03: Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries*, pages 75–86, Washington, DC, USA, 2003. IEEE Computer Society.

[DHM05] Gordon Dahlquist, Brian Hoffman, and David Millman. Integrating digital libraries and electronic publishing in the dart project. In *JCDL '05: Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, pages 114–120, New York, NY, USA, 2005. ACM Press.

[DLF] DLF. Digital libray federantion. Disponível em: http://www.diglib.org. Acesso em: 02 out. 2006.

[dSHL02] Hebert Van de Sompel, Herbert, and Carl Lagoze. Notes from the interoperability front: A progress report on the open archives initiative. *European Conference on Research and Advanced Technology for Digital Libraries*, pages 144–157, 2002.

[dSNLW04] Herbert Van de Sompel, Michael L. Nelson, Carl Lagoze, and Simeon Warner. Resource harvesting within the oai-pmh framework. *D-Lib Magazine*, 10(12), December 2004.

[EN98] Ramez A. Elmasri and Shamkant B. Navathe. *Fundamentals of Database Systems*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1998.

[Fer02] Katia Gomes Ferreira. Teste de usabilidade, 2002. Monografia de Final de Curso. Especialização em Informática. Departamento de Ciência da Computação. Universidade Federal de Minas Gerais.

[FM98] Edward A. Fox and Gary Marchionini. Toward a worldwide digital library. *Commun. ACM*, 41(4):29–32, 1998.

[Fora] Open Archives Forum. History and development of oai-pmh. Disponível em: http://www.oaforum.org/tutorial/english/page2.htm. Acesso em: 12 nov. 2006.

[Forb] Open Archives Forum. Implementing oai-pmh. Disponível em: http://www.oaforum.org/tutorial/english/page4.htm. Acesso em: 12 nov. 2006.

[Forc] Open Archives Forum. Main technical ideas of oai-pmh. Disponível em: http://www.oaforum.org/tutorial/english/page3.htm. Acesso em: 12 nov. 2006.

[Gal05] Marcos Galindo. Experimentando novos modelos de investigação em ciência da informação: O caso liber, 2005. Conferência Universidade do Porto.

[Gar03] Patrícia A. Garcia. Provedores de dados de baixo custo: publicação digital ao alcance de todos. Master's thesis, Universidade Federal do Paraná, Curitiba, 2003. Dissertação (Mestrado em Informática).

[Gro] IETF Working Groups. Rfc 3066. Disponível em: http://www.ietf.org/rfc/rfc3066.txt. Acesso em: 11 abr. 2006.

[Gro05] CDP Metadata Working Group. Dublin core metadata best practices, September 2005.

[GS03] Evan Golub and Ben Shneiderman. Dynamic query visualisations on world wide web clients: a dhtml solution for maps and scattergrams. *International Journal of Web Engineering and Technology (IJWET)*, 1(1), 2003.

[GYa04] Carol J. Godby, Jeffrey A. Young, and Eric Childress and. A repository of metadata crosswalks. *D-Lib Magazine*, 10(12), December 2004.

[HH94] Deborah Hix and H. Rex Hartson. Book review: Developing user interfaces: Ensuring usability through product and process. *SIGCHI Bull.*, 26(1):74–75, 1994. Reviewer-Jenny Preece.

[IBI] IBICT. Biblioteca digital de teses e dissertações. Disponível em: http://bdtd.ibict.br/bdtd/. Acesso em 16 mar. 2006.

[IEE] IEEE. Ieee learning object metadata. Disponível em: http://ltsc.ieee.org/wg12/. Acesso em 15 mar. 2006.

[IS] Ismail Khalil Ibrahim and Wieland Schwinger. Data integration in digital libraries: Approaches and challenges.

[JLHdS04] Henry N. Jerez, Xiaoming Liu, Patrick Hochstenbach, and Herbert Van de Sompel. The multi-faceted use of the oai-pmh in the lanl repository. In *JCDL '04: Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*, pages 11–20, New York, NY, USA, 2004. ACM Press.

[KPT03] Ellen Knutson, Carole Palmer, and Michael Twidale. Tracking metadata use for digital collections. *DCMI Conference*, 2003.

[LdS01] Carl Lagoze and Herbert Van de Sompel. The open archives initiative: building a low-barrier interoperability framework. In *ACM/IEEE Joint Conference on Digital Libraries*, pages 54–62, 2001.

[Lei98] Barry M. Leiner. The ncstrl approach to open architecture. *D-Lib Magazine*, December 1998.

[LGB99] Steve Lawrence, C. Lee Giles, and Kurt Bollacker. Digital libraries and Autonomous Citation Indexing. *IEEE Computer*, 32(6):67–71, 1999.

[LMZN01] Xiaoming Liu, Kurt Maly, Mohammad Zubair, and Michael L. Nelson. Arc - an oai service provider for digital library federation. *D-Lib Magazine*, 7(4), April 2001.

[Lou03] Sebastião Lour. Sistema Único de informação em saúde: Integração dos dados da assistência suplementar à saúde ao sistema sus. *Agência Nacional de Saúde Suplementar*, July 2003.

[LPV04] Marcos G. Lima, Marcos S. Pereira, and Cleiton M. Vieira. Bibliotecas digitais e metadados: uma abordagem integradora. *II Simpósio Internacional de Bibliotecas Digitais*, 2004. Disponível em http://libdigi.unicamp.br/document/?code=8283. Acesso em: 14 out. 2005.

[Lyn97] Clifford A. Lynch. The z39.50 information retrieval standard. part i: A strategic view of its past, present and future. *D-Lib Magazine*, April 1997.

[Men05] Eva Mendez. 10 anos de dublin core. y muchos mas de vocabularios. *Thinkepi*, June 2005. Disponível em: http://www.thinkepi.net/repositorio/10-anos-de-dublin-core-y-muchos-mas-de-vocabularios. Acesso em: 02 abr. 2006.

[MG01] Alexa T. McCray and Marie E. Gallagher. Principles for digital library development. *Commun. ACM*, 44(5):48–54, 2001.

[MMGG01] Marilyn McClelland, David McArthur, Sarah Giersch, and Gary Geisler. Challenges for service providers when importing metadata in digital libraries. *D-Lib Magazine*, 8(4), April 2001.

[MySa] MySQL. Full-text search functions. Disponível em: http://www.mysql.com. Acesso em: 03 out. 2006.

[MySb] MySQL. Mysql:the world's most popular open source database. Disponível em: http://www.mysql.com. Acesso em: 03 out. 2006.

[OAIa] OAI. The open archives initiative protocol for metadata harvesting. Disponível em http://www.openarchives.org/OAI/openarchivesprotocol.html. Acesso em: 12 nov. 2006.

[OAIb] OAI. Página oficial do open archives initiative. Disponível em: http://www.openarchives.org. Acesso em: 12 nov. 2006.

[OAIc] OAI. Registered data providers. Disponível em http://www.openarchives.org/Register/BrowseSites. Acesso em: 03 ago. 2006.

[OAId] OAI. Specification and xml schema for the oai identifier format. Disponível em http://www.openarchives.org/OAI/2.0/guidelines-oai-identifier.htm. Acesso em: 12 nov. 2006.

[OAIe] OAIster. Oaister - about. Disponível em: http://oaister.umdl.umich.edu/o/oaister/. Acesso em: 24 maio 2006.

[Obj] Business Object. O desafio da integração de dados nas grandes corporações. Disponível em: http://www.businessobjects.com.br/. Acesso em 12 mar. 2006.

[oC] Library of Congress. Marc standards. Disponível em: http://www.loc.gov/marc/. Acesso em: 04 out. 2006.

[OCL05] OCLC. *Manual OAICat*, abr 2005. Disponivel em: http://pubserv.oclc.org/oaicat/jars/docs/index.html. Acesso em: 02 fev. 2006.

[PBJ+00] Andreas Paepcke, Robert Brandriff, Greg Janee, Ray Larson, Bertram Ludaescher, Sergey Melnik, and Sriram Raghavan. The z39.50 information retrieval standard. part i: A strategic view of its past, present and future. *D-Lib Magazine*, 6(3), March 2000.

[PH02] Christopher J. Prom and Thomas G. Habing. Using the open archives initiative protocols with ead. In *JCDL '02: Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*, pages 171–180, New York, NY, USA, 2002. ACM Press.

[PHPa] PHP. Php: Hypertext preprocesso. Disponível em: http://www.php.net. Acesso em: 02 out. 2006.

[PHPb] PHP. Simplexml functions. Disponível em: http://www.php.net/simplexml. Acesso em: 10 out. 2006.

[PJ03] Andy Powell and Pete Johnston. Guidelines for implementing dublin core in xml, April 2003. Disponível em: http://dublincore.org/documents/dc-xml-guidelines/. Acesso em: 11 abr. 2006.

[PPG02] Marco Pirri, Maria C. Pettenati, and Dino Giuli. Design of a federation service for digital libraries:the case of historical archives in theporta europa portal (pep) pilot project. *DC-2002: Metadata for e-Communities: Supporting Diversity and Convergence*, October 2002.

[Pre92] Roger S. Pressman. *Software engineering (3rd ed.): a practitioner's approach*. McGraw-Hill, Inc., New York, NY, USA, 1992.

[PS05] Dirk Pieper and Friedrich Summann. Beyond oai services: Bielefeld academic search engine (base). *Joint Workshop on Electronic Publishing*, April 2005.

[RH02] Marcia Rosetto and Adriana H.Nogueira. Aplicação de elementos metadados dublin core para descriçãoo de dados bibliográficos on-line da biblioteca digital de teses da usp. *XII Simpósio Nacional de Bibliotecas Universitárias*, 2002.

[Ros97] Márcia Rosetto. Uso do protocolo z39.50 para recuperação de informação em redes eletrônicas. *Revista Ciência da Informação*, 26(2), 1997.

[Saf95] W. Saffady. Digital library concepts and technologies for the management of library collections : an analysis of methods and costs. *Library technology reports*, 31(3):221–380, 1995.

[Sak05] Angela R. Sakamoto. Informação: O bem mais precioso da empresa. *GNOSIS IT Knowledge Solutions Page*, 2005. Disponível em: http://www.gnosisbr.com.br/artigos/AS0001.html. Acesso em: 04 mar. 2006.

[SKS98] Abraham Silberschatz, Henry F. Korth, and S. Sudershan. *Database System Concepts*. McGraw-Hill, Inc., New York, NY, USA, 1998.

[SL01] Ana C. Salgado and Bernadette F. Lóscio. Integração de dados na web. *XXI Congresso da Sociedade Brasileira de Computação*, 2001.

[Som01] Ian Sommerville. *Software engineering (6th ed.)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2001.

[Sul01] Hussein Suleman. Enforcing interoperability with the open archives initiative repository explorer. In *ACM/IEEE Joint Conference on Digital Libraries*, pages 63–64, 2001.

[Ten01] Roy Tennant. Different paths to interoperability. *Library Journal*, 126(3), February 2001.

[Tro98] Valsoir Tronchin. Modelagem e implementação de data warehouses. *Fenasoft/98*, July 1998.

[Ubi] UbiLib. The simple digital library interoperability protocol (sdlip-core). Disponível em: http://dbpubs.stanford.edu:8091/ testbed/doc2/SDLIP/. Acesso em: 19 mar. 2006.

[WMBB00] Ian H. Witten, Rodger J. McNab, Stefan J. Boddie, and David Bainbridge. Greenstone: A comprehensive open-source digital library software system. In *Proceedings of the Fifth ACM International Conference on Digital Libraries*, 2000.

[wS] w3 Schools. *XML Tutorial*. w3C. Disponível em: http://www.w3schools.com/xml/. Acesso em: 02 abr. 2006.

[ZC06] Marcia L. Zeng and Lois M. Chan. Metadata interoperability and standardization. a study of methodology part i. *D-Lib Magazine*, 12(6), June 2006.

Este volume foi tipografado em LaTeX na classe UFPEThesis (www.cin.ufpe.br/~paguso/ufpethesis).