JOÃO BOSCO DELFINO JÚNIOR

APLICAÇÕES DE ALGORÍTIMOS USANDO VISÕES MATERIALIZADAS PARA INTEGRAÇÃO DE BIBLIOTECAS DIGITAIS

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Informática, Curso de Pós-Graduação em Informática, Setor de Ciências Exatas, Universidade Federal do Paraná.

Orientador: Prof. Dr. Marcos S. Sunye

CURITIBA

2006

Ao meu pai (In Memorian), minha mãe, minha esposa e meus irmãos,

pelo amor infinito.

AGRADECIMENTOS

Em primeiro lugar a Deus, pela força de vontade.

Ao Prof. Dr. Marcos Sunye, pela orientação cedida.

Ao Prof. Dr. Guilherme Ataíde e Profª. Dra. Eliany Alvarenga, pelo incentivo constante.

A Profª. Dra. Maria Salete e Profª. Dra. Laura Sánchez, pela ajuda extra.

Aos funcionários e professores do Departamento de Informática, pela atenção inestimável e pela contribuição.

A todos aqueles que direta ou indiretamente contribuíram para a realização deste trabalho.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

DCMI	-	Dublin Core Metadata Initiative
GPL	-	General Public License
HPL	-	Hewlett Packard Labs
HTTP	-	Hypertext Transport Protocol
IBICT	-	Instituto Brasileiro de Informação em Ciencia e Tecnologia
MIT	-	Massachusetts Institute of Technology
OAI	-	Open Archive Initiative
OAI-PMH	-	Open Archive Initiative Protocol Metadata Harvesting
OJS	-	Open Journal System
PKP	-	Public Knowledge Project
SEER	-	Sistema Eletrônico de Editoração de Revistas
SQL	-	Structured Query Language
UFPR	-	Universidade Federal do Paraná
UNICAMP	-	Universidade Estadual de Campinas
URL	-	Uniform Resource Locator
USP	-	Universidade de São Paulo
VTLS	-	Virginia Tech Library System
XML	-	eXtended Markup Language
XSLT	-	eXtensible Stylesheet

RESUMO

Com a disseminação fluente e constante da informação e através do avanço da informática, foi constatada a necessidade de obter sistemas que manipulem tais informações de forma que atendam à demanda por mais qualidade das mesmas sem interferir na sua integridade. A partir desses fatos foram criadas as Bibliotecas Digitais, com o objetivo de prover recursos, de forma que toda a informação armazenada esteja prontamente e economicamente disponível para o uso por uma comunidade específica ou coleções de comunidades.

Mas para oferecer estes serviços é necessário que as Bibliotecas Digitais estabeleçam uma comunicação entre as mesmas. Com o intuito de gerar esta funcionalidade de interoperabilidade entre as mesmas, foi estabelecido um protocolo chamado *Open Archives Initiative Protocol for Metadata Harvesting*, cuja principal função é colher registros contendo metadados a partir de repositórios (Bibliotecas Digitais), e disponibilizá-los.

A partir do protocolo, os usuários podem usufruir de todas as funcionalidades fornecidas pelas Bibliotecas Digitais, Revistas Eletrônicas e Sistemas de Informação em geral. Mas ainda nos deparamos com um grande problema, que é a falta de integração de diferentes sistemas de disponibilização de conteúdo, localizados em uma determinada instituição de ensino e pesquisa. Isso provoca uma grande perda de tempo e baixo rendimento a nível de produtividade.

Dessa forma esse trabalho modela uma estrutura de integração de conteúdo digital de instituições de ensino e pesquisa, através do uso de visões materializadas. A partir desse mecanismo é possível realizar uma única pesquisa para obter inúmeros resultados em comum, os quais estão em diferentes localidades remotas. Outra grande vantagem é a rápida resposta que tais sistemas fornecem, já que os mesmos estão alocados fisicamente, em diferentes locais situados na Universidade Federal do Paraná e o mecanismo de busca não precisa abranger um âmbito maior no globo, além da própria instituição.

Palavras-Chaves: Biblioteca Digital, OAI, Provedores de Serviços, Conteúdo Digital, Metadados.

ABSTRACT

With the big dissemination and Constant of the information and through the advance of computer science, were evidenced the necessity to get systems that manipulate such information of form that they take care of the demand for more quality of the same ones without intervening with its integrity. To leave of these facts the Digital Libraries had been created, with the objective to provide resources, of form that all the stored information is readily and economically available for the use for a specific community or collections of communities.

But to offer these services it is necessary that the Digital Libraries make a communication between them. With intention to generate this functionality of interoperability between the same ones, a called protocol Open Archives Initiative Protocol for Metadata Harvesting was established, whose main function is spoon registers contends metadata from repositories (Digital Libraries), ant to available them.

From the protocol, the users can usufruct of all the functionalities supplied for the Digital Libraries, Electronic Magazines and Information Systems in general. But we still across a great problem; that it is the lack of integration of different systems to available of content, located in one determined institution of education and research. This provokes a great loss of time and low income the productivity level.

Of this form this work shapes a structure of integration of digital content of education institutions and research, through the use of materialized views. To leave of this mechanism it is possible to carry through an only research to in common get innumerable results, which are in different remote localities. Another great advantage is the fast reply that such systems supply, since the same ones are placed physically, in different situated places in the Federal University of the Paraná and the search mechanism does not need to enclose a bigger scope in the globe, beyond the proper institution.

Key-Words: Digital Librarie, OAI, Service Provider, Digital Content, Metadata

1 INTRODUÇÃO

1.1 MOTIVAÇÃO

Com o avanço da tecnologia e o crescimento da divulgação eletrônica, surgiu o conceito de Bibliotecas Digitais, as quais possuem como principal objetivo, a disponiblização da informação no meio eletrônico para uma determinada comunidade ou conjunto de comunidades [1].

No entanto, para as Bibliotecas Digitais funcionarem sem restrições é preciso que as mesmas forneçam um mecanismo de comunicação. Através dessa ligação elas poderão trocar informações, aumentando assim sua capacidade de suprir todas as necessidades dos usuários. A partir dessa forma de interoperabilidade, as Bibliotecas Digitais poderão alcançar um espaço maior na comunidade científica, se comparado a uma área, onde as mesmas trabalhassem individualmente.

Para que essa integração entre diferentes sistemas de informação seja construída, é necessário que os mesmos trabalhem com o conceito de metadados. Metadados são basicamente definidos como “dados que descrevem dados”, ou seja, dados que possuem informações, as quais auxiliam nos processos de estruturação, localização, descrição, identificação e recuperação de documentos, cujos conteúdos estão no meio digital ou não [2]. As vantagens de se ter sistemas de disponibilização de conteúdos baseados em metadados são inúmeras; podemos citar algumas como: fundação de padrões de dados diante das inúmeras formas de disponibilização de informações na *World Wide Web*, troca das informações armazenadas em sistemas e instituições que utilizam o conceito de metadados aqui descrito, maior exatidão na recuperação das informações cobiçadas, entre diversas outras.

Uma Biblioteca Digital que não está de acordo com as exigências da OAI (Iniciativa de Arquivos Abertos) está praticamente isolada do contexto mundial. A OAI é uma iniciativa para disponibilizar acesso fácil e gratuito aos conteúdos digitais espalhados pela Internet, através da comunicação entre repositórios de conteúdo digital, os quais exercem funções de divulgação, compartilhamento e armazenamento de metadados [3].

A OAI opera através do protocolo OAI-PMH (*Open Archives Initiative Protocol for Metadata Harvesting*), o qual possui a finalidade de realizar a colheita de metadados dos respectivos repositórios de dados, cadastrados na iniciativa. A comunidade ligada à iniciativa é dividida em dois grupos distintos: os Provedores de Serviços (*Service Provider*) e os Provedores de Dados (*Data Provider*). Os Provedores de Serviços oferecem buscas em cima dos metadados armazenados nos repositórios. Já os Provedores de Dados oferecem os respectivos repositórios como ferramenta de armazenamento e disponibilização de todo seu acervo eletrônico [4].

A colheita de metadados e realizada independentemente dos formatos dos dados. O necessário é a consciência de cada comunidade em determinar e trabalhar com formatos que sejam do conhecimento de todos os membros das mesmas. A fins de interoperabilidade básica, o protocolo usa o Dublin Core Unqualified como formato de metadados.

O OAI-PMH opera em cima de alguns verbos responsáveis pela recuperação dos metadados. Cada verbo tem a finalidade de obter uma resposta a uma devida requisição enviada pelo protocolo. Embora o padrão de verbos do OAI-PMH seja o mais utilizado atualmente em todo o contexto mundial para aquisição de informações científicas localizadas remotamente, ainda nos deparamos com o problema da concorrência por tais informações, uma vez que as mesmas estão sendo requisitadas por milhares de usuários espalhados por todo o planeta. Muitas dessas informações estão armazenadas em sistemas de informações alocados em instituições de ensino e pesquisa, onde o acesso dentro das mesmas deveria ser mais rápido e menos competitivo, mas devido à falta de uma estrutura interna de busca integrada, os usuários são obrigados a realizar pesquisas a nível mundial de conteúdos armazenados bem próximo de sua localização física.

Nos deparamos também com um grande problema encontrado em várias instituições de ensino, como a USP (Universidade de São Paulo) e a UNICAMP (Universidade Estadual de Campinas), o qual se concentra na falta de integração dos diversos meios de disponibilização da informação. Meios que vão desde a automatização do acervo das bibliotecas convencionais até a implantação de Bibliotecas Digitais totalmente independentes de material físico encontrados nas

bibliotecas comuns. Com esta falta de homogeneização entre diferentes sistemas, o usuário deve realizar uma mesma consulta diversas vezes, em sistemas específicos, ou seja, o usuário perde tempo e produtividade realizando uma consulta em cada um dos sistemas; onde o mais apropriado seria a existência de uma busca unificada, através da mesma o pesquisador realizaria uma única consulta validada para todos os sistemas, os quais estariam previamente interligados, aumentando assim a interoperabilidade entre os sistemas, proporcionando uma maior facilidade de acesso aos documentos contidos nos mesmos.

Isto é possível, se pensarmos em termos de visões materializadas. De acordo com CRISTINA [5], uma visão é dita materializada quando ela é realmente armazenada na base de dados em vez de ser apenas visualizada a partir das relações base em resposta a consultas. Através das visões materializadas, podemos extrair informação de um dado sistema e por esta mesma informação em outro sistema, realizando assim, uma replicação de dados e tornando ambos os sistemas integrados.

1.2 OBJETIVO

Tendo em vista a necessidade da criação de um modelo de integração de dados, o presente trabalho tem por objetivo modelar uma estrutura de integração de sistemas de informação fazendo uso do conceito de visões materializadas para a construção do mesmo. A partir desta modelagem será possível identificarmos os pontos essenciais para a execução do trabalho em cima de diferentes sistemas de informação e realizar a divulgação do mesmo para instituições e centros de ensino e pesquisa que sofrem com a falta de integração de seus meios de armazenamentos digitais. A efetiva implementação permitiu uma análise detalhada do modelo de integração de dados, a partir do manuseio de metadados, utilizados pelos sistemas de informações oferecidos pela UFPR e de um trabalho em comum com os as tecnologias XML e XSLT. A criação de tal modelo fornece um mecanismo de colheita de metadados mais robusto, uma vez que o mesmo abrange uma área integrada e centralizada a nível institucional.

1.3 METODOLOGIA

A metodologia para o levantamento de informações baseou-se na análise dos sistemas, na bibliografia existente, tal como artigos, manuais e periódicos científicos, e também nas informações constantes das páginas oficiais das entidades citadas.

Para efetiva comprovação das informações aqui apresentadas, foi realizada a integração dos sistemas DSPACE, VIRTUA e SEER localizados no Departamento de Informática da UFPR, onde foi possível validar o que foi levantado, bem como realizar análises e propor a divulgação do trabalho exposto.

O trabalho está estruturado da seguinte maneira: o segundo capítulo aborda o conceito e definições das Bibliotecas Digitais, assim como mitos, desafios, domínio público e acessibilidade das mesmas; no terceiro capítulo encontramos informação sobre a Iniciativa Open Archives, com definições, características, padrões e modelos atuais; o quarto capítulo define a integração de conteúdo abordada, estabelecendo uma análise geral dos sistemas envolvidos; no quinto capítulo mostramos alguns dos algoritmos de visão materializada, assim como a diferença entre eles; e por último relatamos trabalhos futuros, tendências da tecnologia e visões futuras.

2 BIBLIOTECAS DIGITAIS

2.1 INTRODUÇÃO

Bibliotecas Digitais são organizações que disponibilizam recursos, incluindo uma equipe especializada para selecionar, estruturar, oferecer acesso intelectual, interpretar, distribuir, preservar a integridade, certificar a persistência contínua das coleções de trabalhos digitais, de forma que toda a informação armazenada esteja prontamente e economicamente disponível para o uso por uma comunidade específica ou coleções de comunidades [1].

O objetivo principal das Bibliotecas Digitais é a disponibilização constante de todo o seu acervo eletrônico, a partir de instituições de ensino e pesquisa, como Universidades e Centros de Pesquisa Integrada, os quais oferecem uma grande taxa de informação. Antigamente a disponibilização de tais documentos era uma tarefa árdua, devido a vários fatores que vão desde a infra-estrutura, que na época era relativamente básica, até a forma de como oferecer os documentos para os usuários. Mas com o surgimento da Internet esses processos foram sendo automatizados gradativamente e várias Bibliotecas Digitais estão surgindo, com a função principal de por à vista trabalhos científicos, como teses e dissertações [6].

2.2 MITOS E DESAFIOS SOBRE A BIBLIOTECA DIGITAL

De acordo com Rodrigues [9], a Internet é alvo constante de mitos e desafios divulgados pela grande imprensa. Para um melhor esclarecimento detalharemos a seguir os mitos e desafios que repercutem sobre a biblioteca digital:

A Internet é a biblioteca digital. É comum escutarmos que os recursos de informação espalhados pela Internet são uma biblioteca digital. Bibliotecas digitais significam uma diversidade de coisas para diferentes pessoas. Na Internet encontramos alguns tipos de informação, como: textos, imagens, vídeo, áudio e resultados de suas combinações.

A grande pergunta que os autores da literatura científica fazem é se esta seqüência de objetos eletrônicos podem ser considerados uma biblioteca digital, de acordo com os padrões tradicionais descritos por uma biblioteca convencional.

Com isto é necessário o desenvolvimento de um mecanismo para a exploração de recursos em rede, através de uma distribuição da informação descentralizada, onde a informação eletrônica distinta seja criada com autonomia. É importante ressaltarmos que esta infra-estrutura necessitará de profissionais qualificados para o manuseio da informação [7].

A realidade de uma única biblioteca digital. Esta forma de pensar faz com que se construam desafios para quebrar as regras tradicionais de licenciamento e de direito de cópia. Tais mudanças afetarão drasticamente os serviços disponibilizados por repositórios digitais, bem como os custos de proibição de disponibilização no meio digital, além do suporte para a infra-estrutura técnica. Já podemos observar diversos repositórios de dados competindo entre si, de modo desorganizado, sendo a biblioteca digital somente uma dentre diversas fontes de informação.

De acordo com esta competitividade, não é difícil pensarmos em coleções digitais privatizadas, acessíveis somente por assinatura. O grande desafio é a construção de normas de interoperabilidade para consultas nesses novos mecanismos de informação [7].

Bibliotecas digitais fornecerão acesso mais igualitário em qualquer lugar e a qualquer tempo. Hoje sofremos com a falta de acesso à Internet em lugares pobres, que não possuem condições financeiras para implantar e manter uma Internet com velocidade aceitável no mercado. Com isso, é imprescindível a disponibilização de acesso universal às redes de telefonia para que a Internet seja o principal meio de envio e recebimento de informações.

Apesar dos grandes esforços realizados pelos governos as previsões para uma Internet mais igualitária esta distante da realidade atual, devido a obstáculos técnicos e legais envolvidos. A definição quanto à legislação de direito autoral é um processo lento e tem o potencial de comprometer a concepção ideal da biblioteca digital, além destes fatores, a administração da tecnologia para bibliotecas digitais está tornando-se

mais complexa que a gerência das licenças e acesso de uso, e neste sentido, o impacto sobre o acesso igualitário poderá ser considerável [7].

Bibliotecas digitais serão mais baratas que bibliotecas impressas. Apesar de existirem muitos projetos de bibliotecas que barateam recursos humanos e financeiros, a idéia de que a biblioteca digital é mais econômica que a biblioteca impressa esta longe de ser estabelecida, já que em alguns casos, a troca para periódicos eletrônicos poderá economizar o recurso orçamentário da biblioteca, repassando o custo de impressão para os usuários.

Hoje temos muitas bibliotecas que investem bastante na infra-estrutura digital, dando assim um crescimento contínuo a novas versões, licenças, atualizações, infra-estrutura administrativa e treinamentos.

Talvez o melhor de todos os mundos possíveis seria uma "biblioteca digital larga, inclusiva" preenchida com uma multidão de objetos e de software informativos de todos os tipos. Tal lugar seria construído da base (com a mais alta tecnologia em hardware) até o topo (com software de última geração), e consistiria em materiais e objetos de bibliotecas tradicionais (e poderia ter recursos para manter). Seria uma versão evoluída do que a Internet é hoje [7].

2.3 DOMÍNIO PÚBLICO E DIREITOS AUTORAIS

Rodrigues [9] relata o grande número de acessos a publicações disponíveis gratuitamente na Internet, onde reparamos problemas históricos ligados ao acesso à informação por parte do domínio público.

Abaixo temos alguns fatores que legalizam o objeto público [9]:

- Hoje temos vários documentos legais arquivados em instituições governamentais e acadêmicas que estão disponíveis para poucos, mas a partir de projetos de digitalização dos mesmos, tais documentos podem abranger uma área maior de disponibilização de acessos, através da Internet;
- Através da disposição de documentos em rede, podemos baratear os custos de manutenção dos mesmos;

- Trabalhos colaborativos que compartilham interesses comuns para produção de materiais digitais, a nível global.

Através da Internet a produção e disposição de documentos é facilitada, criando assim novos locais públicos constituídos por informações gratuitas ou a um custo muito baixo. Mas para tal mecanismo é necessário implantar padrões de confiabilidade para o domínio público.

Para atender a esta demanda de padrões de confiabilidade, é adotado o mecanismo conhecido como *copyleft*. Tal mecanismo é definido como um modelo de licença pública para programas de computadores, conhecido também como GPL *General Public License*. Esta licença permite ao usuário realizar cópias e modificações nos documentos, mas o proíbe de solicitar direitos autorais sobre o material modificado.

O termo *copyleft* vem de um trocadilho em inglês, que substitui o "*right*" (direita, em inglês) de *copyright* por "*left"*, (esquerda, em inglês). O duplo sentido do termo está no fato de que a palavra "*left"* é o verbo "*leave"* (deixar) no passado, tornando *copyleft* um termo próximo a "cópia autorizada" [8].

No caso do termo *copyleft*, os usuários possuem permissão do próprio autor para copiar, adicionar, ou modificar um produto original; ao contrário da frase "todos os direitos reservados" onde é permitido a cópia parcial ou total do objeto e sua disposição na Internet apenas para uso pessoal, sem fins lucrativos.

2.4 ACESSIBILIDADE DE DOCUMENTOS DIGITAIS

Deparamos com um grave problema de incompatibilidade de formatos que necessitam de modificações e transferências e freqüentemente ocorre perda da estrutura e conteúdo. Isso se da devido ao fato de os editores e autores não estarem sincronizados com a revolução das tecnologias da informação [9].

De uma forma geral um documento digital, além de depender de um programa específico, depende também de uma seqüência de equipamentos. Vários fatores devem

ser padronizados para assegurar que determinados documentos digitais sejam acessados no futuro.

Uma das formas de tornar os documentos digitais continuamente acessíveis é realizando migrações para formatos mais atuais, como os discos de vinil, transferidos para CD-ROMs e atualmente para DVDs. Mas se tal migração não for realizada muitos documentos digitais serão perdidos para sempre [9].

De acordo com Rodrigues [9], existem duas estratégias de preservação de documentos digitais. A primeira é construir um sistema independente para comportar uma versão específica destes documentos. A segunda seria prolongar a vida útil dos sistemas atuais para que os documentos se tornem acessíveis usando seu programa original.

Mas apesar dos esforços, a experiência mostra que sistemas padronizados não são viáveis, pois os mesmos não acompanham a rápida evolução das tecnologias da informação.

O problema é que, cada migração de documentos digitais para novos formatos tecnológicos, acarreta a perda de informação, tornando assim a versão do documento final diferente da inicial.

Para criar padrões de preservação dos documentos digitais Rodrigues [9] sugere a construção de normas para codificações, permitindo a interpretação de documentos digitais, os quais estão em formatos não padronizados, possibilitando assim uma futura acessibilidade aos documentos através de emulações do ambiente em que o documento digital foi produzido.

Dessa forma, devemos associar aos documentos digitais informações que indiquem sua procedência e anotações que possibilite a transferência dos mesmos para futuros padrões criados por novas tecnologias. Com isso podemos afirmar, que para permitir que um documento digital ultrapassado possa ser acessado futuramente é necessário uma migração contínua dos documentos digitais para as novas formas de disposição asseguradas pela tecnologia crescente, procurando sempre preservar o documento original, garantindo assim a autenticidade da informação a qual ele representa.

3 A INICIATIVA OPEN ARCHIVES

3.1 INTRODUÇÃO

A *Open Archives Initiative* desenvolve e promove padrões de interoperabilidade, para facilitar a disseminação eficiente do conteúdo. A OAI também tem como objetivo realçar o acesso aos diversos repositórios de dados cadastrados na mesma, com o intuito de aumentar a disponibilidade da comunicação acadêmica. O suporte contínuo deste trabalho fortalece um trabalho conjunto da OAI [4].

Os padrões estipulados pela OAI são de suma importância para desenvolver a interoperabilidade e disponibilizar acesso aos mais diversos trabalhos armazenados na Internet, os quais utilizam este padrão adotado. É importante ressaltarmos que os protocolos construídos são independentes dos conteúdos disponibilizados e de questões financeiras que possam afetar o acesso aos trabalhos armazenados.

Deparamos com termos como "*Archive*" que significa local para armazenamentos de conteúdo digital [10] e "*Open*", que na língua portuguesa significa aberto. Isto não quer dizer que possuirmos acesso gratuito aos repositórios participantes da Iniciativa, ou seja, apenas o protocolo disponibilizado pela *Open Archives* é aberto ao público.

A OAI é formada por um comitê executivo, o qual controla os detalhes operacionais da iniciativa sob a direção do comitê principal e com o suporte do comitê técnico, cuja principal função é disponibilizar uma coleção de comitês técnicos, organizados por objetivos específicos como a disponibilização do protocolo. Estes comitês avaliam a eficácia da arquitetura de interoperabilidade OAI e propõem mudanças e reformulações baseadas na experiência da comunidade. A OAI é patrocinada pela *Digital Library Federation, Coalition for Networked Information* e *Natural Science Foundation* [11].

A Iniciativa tem como base o protocolo Open Archives Initiative Protocol for Metadata Harvesting. Tal protocolo faz com os membros participantes da Iniciativa compartilhem seus metadados com os demais. Metadados são basicamente definidos como "dados que descrevem dados", ou seja, dados que possuem informações, as

quais auxiliam nos processos de estruturação, localização, descrição, identificação e recuperação de documentos, cujos conteúdos estão no meio digital ou não [2].

Os membros da OAI formam os Provedores de Serviços e Provedores de Dados. Os provedores de Serviços oferecem buscas em cima dos metadados armazenados nos repositórios. Já os provedores de dados oferecem os respectivos repositórios como ferramenta de armazenamento e disponibilização de todo seu acervo eletrônico. Em termos gerais podemos citar as Bibliotecas Digitais como Provedores de Dados e ferramentas de buscas específicas na área científica como Provedores de Serviços.

Através do cadastramento das instituições no *site* da OAI, as mesmas podem operar como Provedores de Serviços ou Dados, fazendo assim, uso de toda a infra-estrutura e suporte disponibilizados pela iniciativa.

É importante informarmos que o *Dublin Core* é o padrão de metadados estipulado pelo protocolo OAI-PMH.

3.2 DUBLIN CORE METADATA INITIATIVE (DCMI)

A Iniciativa de Metadados Dublin Core (DCMI) é uma organização dedicada a promover a adoção difundida de padrões de metadados para interoperabilidade e desenvolvimento de vocabulários especializados em metadados para a descrição de recursos, permitindo a construção de sistemas de informação mais inteligentes. O objetivo principal deste padrão é tornar fácil a busca de recursos usando a tenologia da *Internet* [12]. A seguir temos a tabela 1, a qual descreve resumidamente os quinze elementos do padrão Dublin Core [13]:

Elemento	Descrição
Title	Nome dado ao recurso.
Subject	Tópico relacionado ao conteúdo do recurso.
Description	Uma breve descrição do conteúdo do recurso.
Type	A natureza ou gênero do conteúdo do recurso.
Source	A referência ao recurso, o qual o presente recurso é derivado.
Relation	Uma referência a um recurso relacionado.
Coverage	Características de extensão ou do escopo do recurso.
Creator	Uma entidade primária responsável pela construção do conteúdo do recurso.
Publisher	A entidade responsável por disponibilizar o recurso.
Contributor	Qualquer entidade responsável por fazer contribuições ao conteúdo do recurso.
Rights	Informação sobre os direitos autorais em cima do recurso.
Format	A manifestação física ou digital do recurso.
Date	A data associada à criação ou alteração do recurso.
Identifier	Um identificador único fornecido ao recurso
Language	A língua do conteúdo intelectual do recurso.

Tabela 1 - Quinze Elementos do Dublin Core.

O Dublin Core é utilizado na complementação de funcionalidades já existentes, no que se refere à pesquisa e indexação de metadados baseados na tecnologia *Web*. No início a principal preocupação era a geração de metadados para recursos eletrônicos. Mas a partir de discussões realizadas em cima dessa meta, os membros da comunidade Dublin Core chegaram a uma conformidade no que diz respeito ao objetivo primordial da utilização dos metadados no padrão Dublin core, ou seja, os mesmos devem ser usados para representar objetos físicos também e não apenas objetos digitais. A partir

dessa idéia os membros da DCMI trabalham em conjunto na disponibilização de projetos, os quais usam o Dublin Core como ferramenta de padronização dos mesmos.

Além dos quinze elementos Dublin Core descritos, nos deparamos com mecanismos nomeados de: “qualificadores Dublin Core”. Tal estrutura tem a finalidade de realizar um aperfeiçoamento na exatidão dos metadados construída pelas aplicações, através de melhoras geradas ao significado de um recurso. Mas a partir deste refinamento, pode ser gerado um alto grau de complexidade, o que lesa significativamente o trabalho dos metadados no que diz respeito à interoperabilidade e à compatibilidade dos mesmos quando usados em sistemas, os quais adotam o padrão Dublin Core. Um exemplo prático deste problema é o elemento “date”. Tal elemento poderia ser facilmente aperfeiçoado para atender a exigências mais específicas, como um tipo exclusivo de data, ou seja, data da última modificação, data de publicação, etc.

3.3 PROVEDORES DE SERVIÇOS

Através dos Provedores de Serviços requisições em cima do protocolo OAI-PMH são construídas e enviadas aos Provedores de Dados, os quais irão fornecer os metadados exigidos.

Os Provedores de Serviços são responsáveis pela colheita, estruturação e disponibilização dos metadados. Devido a um imenso assentimento à iniciativa por parte da comunidade científica global, as respostas de requisições realizadas através dos Provedores de Serviços estão a cada dia tornando-se mais precisas e numerosas. Os Provedores de Serviços são registrados pelos membros participantes da OAI no *site* da *Open Archives* [14].

Podemos citar como exemplo de um mecanismo provedor de serviços o OAISTER, o qual foi construído e implantado pela *University of Michigan Digital Library Production Services*. Através dessa ferramenta é possível ter acesso a mais de 160 Provedores de Dados devidamente cadastrados na OAI, a partir de um único ponto.

A partir deste provedor de serviços temos diversas maneiras de realizar buscas, como: consultas efetuadas através de frases, palavras-chave, autores e títulos de

trabalhos; os quais serão procurados em todo o corpo dos metadados, ou apenas em campos específicos, como: título e autor.

3.4 PROVEDORES DE DADOS

Provedores de Dados são repositórios de conteúdo eletrônico, os quais são exportados, utilizando as regras da OAI. Os Provedores de Dados exportam os metadados dos seus registros, sem a necessidade de expor o seu conteúdo por completo. Muitos Provedores de Dados incluem um campo de metadado adicional, contendo a URL (*Uniform Resource Locator*), que direciona o usuário para o texto completo descrito pelos metadados. É importante ressaltarmos que a função do protocolo OAI-PMH é colher os metadados, ou seja, o protocolo OAI – PMH não é responsável pela ligação contida entre o metadado colhido e o seu conteúdo por inteiro.

Alguns dos provedores de dados são criados com suporte ao protocolo OAI-PMH, através da implementação de programas que apóiam a Iniciativa; outros precisam construir uma interface com o protocolo, por serem bases de dados pré-existentes.

A OAI oferece duas grandes vantagens para os membros participantes: a primeira, de realizar a adoção de um padrão que está cada vez mais se consolidando na comunidade científica para a disseminação da informação; e a segunda, de aumentar os acessos aos trabalhos disponibilizados pelas instituições e centros de ensino e pesquisa de forma barata e eficiente.

Apesar de termos uma grande variedade de repositórios compatíveis com a OAI, muitos ainda não constam na lista da página da Iniciativa [14], a maioria deles por não terem migrado para a mais nova versão 2.0 do OAI-PMH porém os mesmos podem ser vistos no provedor de serviços OAIster [15].

No *site* da iniciativa temos alguns provedores de dados devidamente registrados na OAI, os quais fornecem seus metadados para colhetas realizadas por sistemas que adotam o protocolo OAI-PMH.

3.5 O PROTOCOLO OAI-PMH

3.5.1 Introdução

O protocolo OAI-PMH foi desenvolvido com o objetivo de oferecer facilidade e eficácia no processo de integração de buscas em cima de bases de dados de pesquisa. Através das diversas funcionalidades disponibilizadas pelo protocolo é possível obter respostas às consultas previamente realizadas com um expressivo grau de precisão, e dessa forma reduzir drasticamente o tempo de retorno destas buscas, já que as informações estão compartilhadas entre os diversos membros da iniciativa.

O formato dos metadados colhidos deve estar em conformidade com os padrões exigidos por comunidades específicas ou conjunto de comunidades que formam um complexo composto por um número determinado de provedores de serviços e dados. Para fins de interoperabilidade, podemos citar como exemplo de formato de metadados o Dublin Core.

Através do protocolo OAI-PMH, temos toda uma estrutura para a realização da colheita de metadados dos repositórios já cadastrados na iniciativa. Dessa forma os Provedores de Dados possuem uma forma simples e eficaz de oferecer suas informações, fazendo uso de tecnologias estáveis, como HTTP (*Hypertext Transport Protocol*) e XML (*eXtended Markup Language*).

3.5.2 Características

O Protocolo OAI-PMH é composto por 6 requisições (*requests*) ou verbos. Todas as respostas a tais requisições são realizadas através de código XML.

3.5.2.1 A linguagem XML

O XML (*eXtended Markup Language*) foi divulgado pelo W3C (*World Wide Web Consortium*) [16] e recomendado a partir de 1998.

A linguagem XML tem a função de definir domínios ou linguagens específicas de um domínio, por isso ela é considerada uma meta linguagem, ou seja, ela exprime dados em um determinado formato [17].

A linguagem XML tem a característica de ser independente de meio de apresentação, pois descreve apenas dados, sendo um simples arquivo texto, ou seja , não necessita de ferramentas visuais e independe de plataforma [18].

3.5.2.2 Harvesting

O protocolo OAI-PMH habilita a definição de harvesting, para realizar colheitas a partir dos repositórios de dados cadastrados na Iniciativa. Através dos Provedores de Serviços os dados coletados periodicamente pelo protocolo são exibidos aos usuários. Na figura 1 temos um modelo desta estrutura unilateral.

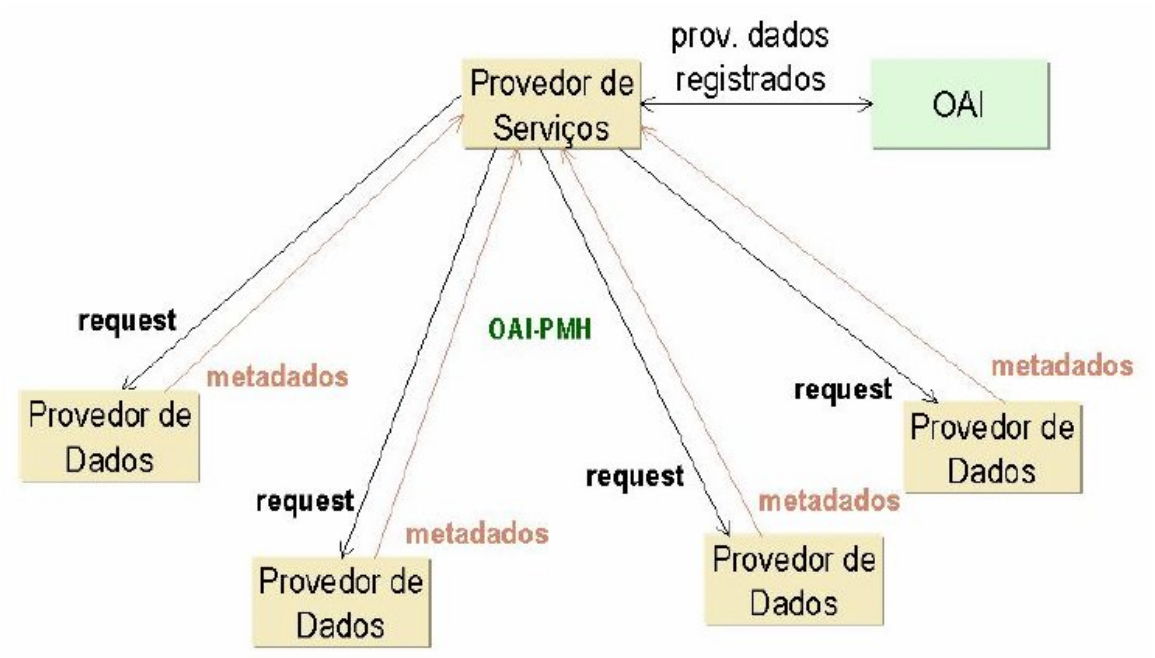

Figura 1 - Harvesting

A coleta de metadados esta baseada em 2 critérios. Baseada em data, ou seja, são colhidos apenas os metadados incluídos ou modificados em uma data específica. Baseada em conjuntos, o protocolo define uma coleção como uma estrutura de agrupamento de itens com o intuito de termos uma coleta seletiva de registros [19].

3.5.2.3 Verbos do OAI –PMH

O protocolo OAI – PMH é composto por seis verbos, os quais são responsáveis pelas solicitações (*requests*) e suas respectivas respostas em XML, ou seja, o provedor de serviços envia uma determinada solicitação ao provedor de dados, e o mesmo responde à devida solicitação.

Identify

Este verbo é utilizado para obter informações específicas sobre o Provedor de Dados, como: nome do provedor de dados, enderêço do repositório (URL), versão do protocolo implementada e endereço eletrônico (*e-mail*) do administrador do repositório.

ListMetadataFormats

Este verbo recupera os formatos de metadados disponíveis no repositório. Podemos definir um identificador de um registro em particular como parâmetro, realizada esta tarefa, os metadados deste determinado registro serão devidamente listados. Vale ressaltar que o formato padrão de metadados é o *Dublin Core* (representado no protocolo por *oai_dc*).

GetRecord

Obtém um único registro do repositório. Devemos determinar o formato dos metadados (*metadataPrefix*), e o identificador do registro (*Identifier*), o qual é único dentre a comunidade OAI.

ListRecords

Este verbo é responsável pela tarefa de colheita dos metadados dos repositórios. Podemos determinar o tipo de colheita a ser realizada, através de argumentos adicionais. Temos dois tipos de colheita, as quais podem ser requisitadas: a coleta seletiva baseada em data (*date-based*) ou a coleta fundamentada em conjuntos (*set-*

based). É importante citarmos a obrigatoriedade da especificação do prefixo do metadado (*metadataPrefix*), cujo padrão é o *oai_dc*, já citado anteriormente.

ListIdentifiers

O verbo *ListIdentifiers* é uma versão resumida do verbo *ListRecords*, que recupera apenas os cabeçalhos dos registros. Nele também podemos determinar o tipo de coleta a ser executada; a coleta seletiva baseada em data (*date-based*) ou em conjuntos (*setbased*). Novamente é importante ressaltarmos o fato obrigatório de especificar o prefixo do metadado (*metadataPrefix*).

ListSets

Este verbo exibi a composição do conjunto de um dado repositório, muito útil para o *harvesting* seletivo baseado em conjuntos.

3.6 O FUTURO DA OAI

O OAI-PMH ainda se depara com problemas, como: o fato de usuários desejarem realizar buscas em lugares específicos, não precisando abranger todo o globo, dessa forma a resposta seria bem mais rápida; a duplicidade das informações extraídas, uma vez que para uma consulta, podemos obter diversas respostas iguais provindas de lugares diferentes; e o problema de sincronização de dados, ou seja, a cópia do metadado fornecido pelo provedor de serviços não combina com o metadado armazenado no devido repositório [31]. Mas apesar disto, esperamos que as bibliotecas digitais que estão surgindo amplamente possuam o protocolo OAI-PMH já acoplado a sua infra-estrutura. Esta idéia vem se consolidando cada vez mais, através de discussões realizadas entre as diversas comunidades que fazem parte do escopo da OAI.

Hoje os comitês organizadores discutem estratégias para a manutenção e evolução do sucesso do protocolo OAI-PMH [20].

4 INTEGRAÇÃO DE CONTEÚDO

4.1 INTRODUÇÃO

Situados na UFPR (Universidade Federal do Paraná) consideramos três dos principais sistemas de informação implantados na mesma. O primeiro deles, o VIRTUA [21] fornece toda uma infra-estrutura de automatização de bibliotecas convencionais. O segundo nomeado DSPACE [22], possui módulos e características de uma Biblioteca Digital. E por último temos o SEER [23] (Sistema Eletrônico de Editoração de Revistas), responsável pela implantação e manutenção das revistas científicas da UFPR.

4.2 VIRTUA

O sistema VIRTUA é disponibilizado a partir da VTLS (*Virginia Tech Library System*), que tem como principal função a construção de **software** dedicado às mais específicas exigências geradas na área de bibliotecas em geral. O VIRTUA oferece algumas características, como: ambiente de pesquisa, integração total com demais bibliotecas, criação de fórums para discussões internas dos usuários da ferramenta, compatibilidade com o Banco de Dados Oracle, disponibilização da informação em diversas línguas e demais vantagens de igual importância [24].

4.3 DSPACE

O DSpace é uma ferramenta implementada pelo MIT (*Massachusetts Institute of Technology*) em parceria com o HPL (*Hewlett-Packard Labs*) e tem como principal função oferecer um repositório de pesquisa digital e produção de material educacional pelos membros de universidades, centros de pesquisa e organizações. O Sistema DSpace provê uma forma de gerenciar materiais de pesquisa e publicações em um repositório profissional, para disponibilizar os mesmos aos usuários constantemente. O

sistema suporta todas funções que uma organização de pesquisa necessita para executar um serviço de repositório digital [25].

4.4 SEER

O SEER é uma adaptação, traduzida e customizada pelo IBICT (Instituto Brasileiro de Informação em Ciência e Tecnologia) baseado no software OJS (*Open Journal System*) desenvolvido pelo PKP (*Public Knowledge Project*) da Universidade British Columbia, seguindo a filosofia *Open Archives*. O SEER é um software desenvolvido para a construção e gestão de uma publicação periódica eletrônica. A ferramenta permiti ações essenciais à automação das atividades de editoração de periódicos científicos [26].

4.5 INTEGRAÇÃO DE SISTEMAS HETEROGÊNEOS

Os sistemas da UFPR considerados neste trabalho estão divididos em três módulos principais. Primeiro temos o SEER, responsável pelas 28 revistas eletrônicas da UFPR; em seguida encontramos o DSPACE, o qual armazena todo o acervo de teses e dissertações, e por último temos o VIRTUA, responsável por guardar todo o acervo de conteúdo digital disponibilizado pelas bibliotecas da UFPR.

A partir de uma busca unificada realizada no PI (Portal da Informação) [32], é possível obtermos várias respostas de sistemas de informações diferentes, ou seja, a partir de uma única consulta podemos receber respostas específicas armazenadas em sistemas localizados em diferentes setores da UFPR.

Na figura 2, podemos visualizar um exemplo de resposta a uma consulta, onde as informações estão em sistemas diferentes.

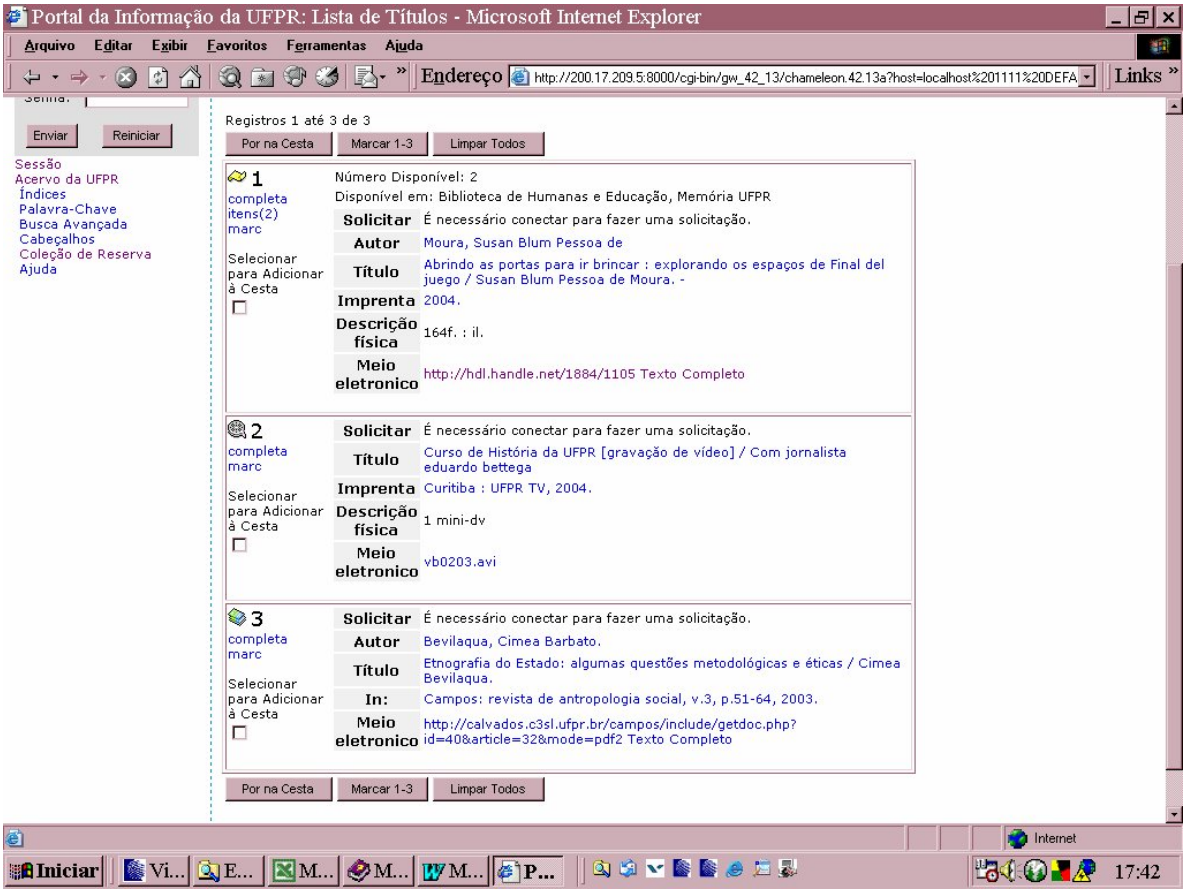

Figura 2 – Resposta à Consulta

Como podemos observar a resposta retornou para uma mesma consulta, três resultados diferentes. O primeiro deles, uma dissertação situada na biblioteca digital DSPACE, o segundo está relacionado a um vídeo encontrado no VIRTUA, e o último refere-se a um artigo armazenado no SEER, relacionado à revista Campos – Revista de Antropologia Social.

Podemos começar pelo VIRTUA, onde o mesmo replica todos os dados do SEER, para poder se comunicar com o mesmo; em seguida temos o DSPACE, o qual realiza uma cópia dos dados referentes a teses e dissertações encontrados no VIRTUA.

Dessa forma podemos afirmar que os sistemas realizam replicação das informações para poderem construir a estrutura de integração entre eles, onde o mais viável seria propor a construção de um modelo formal de integração de diferentes sistemas de informação, através da execução de um algoritmo específico.

É importante ressaltar que a estrutura de integração corrente é simplesmente uma cópia dos metadados dos sistemas envolvidos sem o cuidado de termos dados duplicados e dados sincronizados.

Na figura 3, temos uma visão de toda a estrutura de integração.

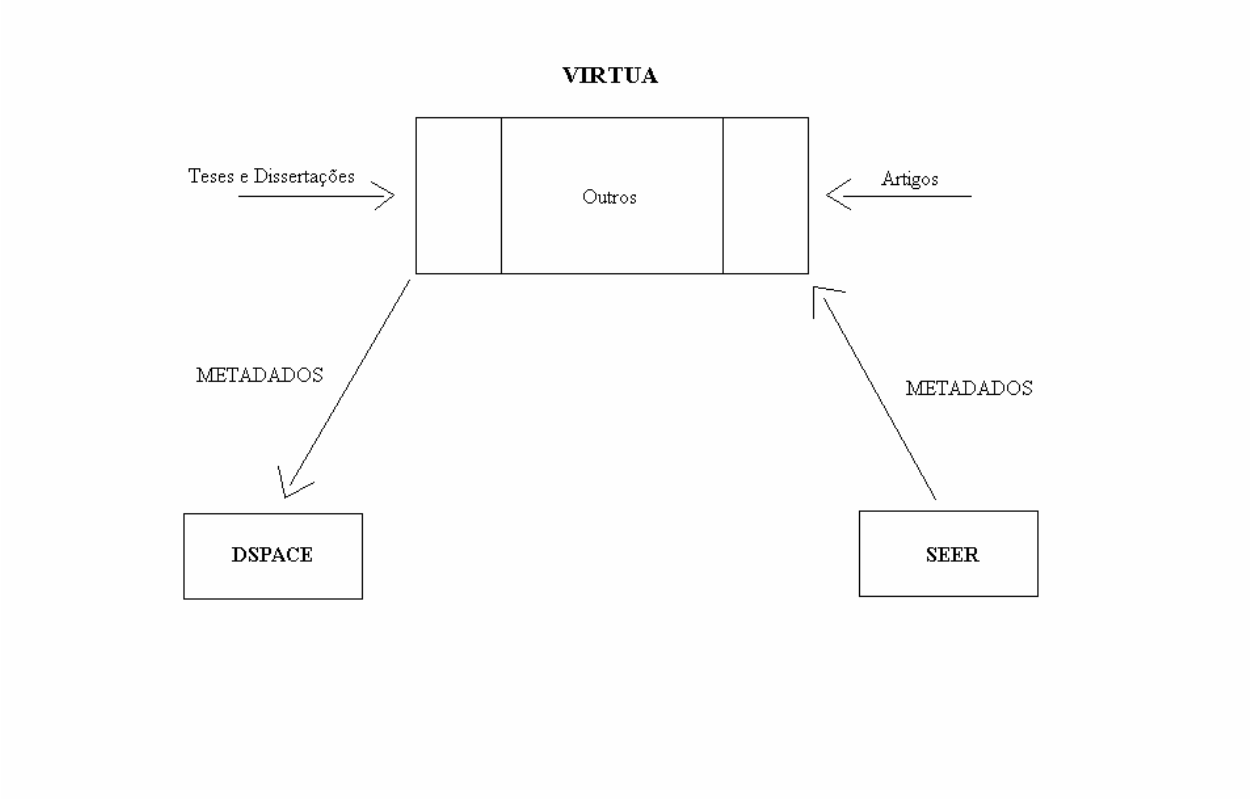

Figura 3 – Estrutura de Integração Atual

A partir desse modelo atual percebemos a necessidade da implantação de um algoritmo consolidado cientificamente, o qual execute a replicação dos metadados de forma contínua e padronizada. Dessa maneira poderíamos encapsular toda a estrutura, construindo assim um modelo formal de integração de sistemas de informações heterogêneos.

5 VISÕES MATERIALIZADAS

5.1 INTRODUÇÃO

Uma visão materializada pode ser vista como um cachê, uma cópia dos dados que pode ser acessada rapidamente [27].

Uma visão é dita materializada quando ela é realmente armazenada na base de dados em vez de ser computada a partir das relações base em resposta a consultas [5].

Uma visão materializada é uma consulta cujo resultado já esta computado e armazenado na base de dados, por isso é uma das principais opções para o controle do desempenho de um Data Warehouse. Através das visões materializadas é possível realizar uma consulta simples, já que as mesmas são similares a uma consulta armazenada em uma tabela. Isto traz ganho de desempenho, uma vez que o resultado final já está previamente armazenado, dispensando assim a execução constante de uma visão específica [28].

A materialização de visões pode ser composta a partir do zero ou mantida incrementalmente por propagação de mudanças de dados do banco de dados para a visão, ou seja, para que a visão possa estar constantemente sincronizada aos dados base, será necessário recriar a visão a partir dos dados de origem ou então, atualizá-la de forma incremental [29].

As visões materializadas são pré-calculadas para melhorar o desempenho das consultas em cima da base de dados [30]. Uma visão materializada pode ser usada para isolar os dados de interesse, permitindo que consultas subseqüentes sejam executadas sobre um conjunto de dados menor e possivelmente mais estruturado que os originais, evitando-se a extração repetitiva de informações já extraídas anteriormente.

Visões definidas sobre fontes de dados que sofrem atualizações pouco freqüentes podem ser materializadas em uma nova fonte de dados, garantindo melhoras de desempenho e maior facilidade no acesso às informações. Desse modo, consultas podem ser submetidas diretamente sobre a visão materializada, evitando repetições contínuas do processo de extração nas fontes, sempre que possível.

5.2 ALGORITMOS DE VISÃO MATERIALIZADA

De acordo com [5], podemos descrever dois algoritmos para manutenção incremental em cima de visões materializadas. Ambos os algoritmos aplicam-se nas transações onde reflete alterações na base de dados (inserções, eliminações, e atualizações); sendo que as visões podem utilizar uniões, negações, agregações e recursões. Os dois algoritmos geram regras que permitem implementar alterações nas visões, a partir das alterações ocorridas nas relações base e a situação das visões materializadas antes das modificações.

O primeiro algoritmo chamado algoritmo de ***counting***, é utilizado em visões cuja mesmas possuem ou não linhas duplicadas. A idéia principal é manter um contador central do número de derivações para cada linha de registro da visão. A seguir veremos um exemplo deste algoritmo para podermos compreender melhor o mesmo.

Exemplo 1:

Dada a relação link = {(a,b), (b,c), (b,e), (a,d), (d,c)} e, como resultado, a visão hop = {(a,c), (a,e)}. A tupla hop (a,e) tem apenas uma única ocorrência, pois é derivada uma vez de link, já a tupla hop (a,c) possui duas derivações. Se a visão não for definida com o operador ***distinct***, ou seja, possuir duplicidade, então teremos ***count*** igual a 1 para hop (a,e) e ***count*** igual a 2 para hop (a,c). Como vemos o algorítmo de ***counting*** relata a duplicidade na visão e armazena a mesma em contadores.

Agora vamos supor que a tupla link (a,b) seja excluída. A visão hop seria executada novamente, e teríamos como resultado hop = {(a,c)}. De acordo com o algoritmo de ***counting*** uma derivação de cada uma das tuplas hop (a,c) e hop (a,e) precisa ser excluída. O algoritmo relata também, baseando-se nos contadores salvos que, a tupla hop (a,c) ainda possui uma derivação, portanto hop (a,e) deve ser eliminada, pois não possui mais derivações.

O segundo algoritmo denominado **DRed** (*Deletion and Rederivation*) e descrito em [5], não pode ser usado em visões, cujas mesmas não possuam o operador ***distinct***, ou seja, as visões não podem obter duplicidade de tuplas. O algoritmo faz um cálculo nas alterações das visões dividido em três etapas. Na primeira etapa, o algoritmo distingue as tuplas derivadas, as quais serão eliminadas, a partir de uma estimativa. Uma tupla *t* se encontra nesta estimativa se qualquer alteração realizada na relação base anular a ocorrência de *t*. A segunda etapa consiste na remoção das tuplas que possuem derivação alternativa na relação base. Por último, as novas tuplas a serem armazenadas são calculadas, a partir das novas inserções executadas na relação base e na construção de uma visão materializada atualizada, conforme a relação base. Para entender melhor o algoritmo, vamos seguir o exemplo abaixo:

Exemplo 2:

Dada a relação link = {(a,b), (b,c), (b,e), (a,d), (d,c)} e, como resultado, a visão hop = {(a,c), (a,e)}. Suponha que eliminássemos a tupla link (a,b). O algorítmo eliminará as tuplas hop (a,c) e hop (a,e), pois ambas dependem da tupla link (a,b), a qual foi eliminada. Após essa etapa, o algoritmo executará uma busca por derivações alternativas de cada uma das tuplas eliminadas. Assim no segundo passo, teremos a nova derivação da tupla hop (a,c) e a inserção da mesma na visão materializada. A terceira etapa do algoritmo é vazia para este exemplo, já que não temos tuplas a serem inseridas na relação link.

A partir dos algoritmos apresentados, podemos observar a relevância dos mesmos para integração dos sistemas da UFPR, ou seja, se comparássemos a solução atual com os exemplos citados, teríamos como relação link: o sistema SEER e as teses e dissertações encontradas no VIRTUA. E como visão hop: os artigos exportados do SEER e o DSPACE (repositório de teses e dissertações).

Conforme [5], o trabalho de manutenção das visões depende fortemente da quantidade de informação manuseada. Inúmeros projetos foram executados no intuito de ilustrar algoritmos que executem o trabalho de manutenção das visões, dependendo da quantidade de informação e do tipo da visão.

Foram estudados casos onde toda a informação está disponível para o processo de manutenção, ou seja, todas as relações base e visões materializadas estão livres para manutenção. Tais estudos levam em consideração características no processo de formação da visão. A seguir podemos observar uma breve descrição de alguns algoritmos, organizados pelos tipos de visões.

1. **Visões não recursivas:** neste caso as técnicas mais apropriadas incluem o algoritmo de ***counting*** (descrito anteriormente) e demais algorítmos que também utilizam-se de contadores na sua estrutura principal. Podemos citar mecanismos que possuem ponteiros, os quais partem da tupla de origem para as tuplas derivadas e também estruturas que definem regras de produção para manter as visões SQL definidas com o operador ***distinct***, agregações e visões onde os atributos das mesmas define o funcionamento da relação base, a qual está sendo atualizada.

2. **Visões com *outer-join*:** nesta classificação temos o algoritmo, o qual redifine toda a visão, obtendo duas únicas instruções (uma instrução com *left-outer-join* e outra com *right-outer-join*) para executar as alterações futuramente implementadas.

3. **Visões recursivas:** podemos citar o algoritmo **DRed** (descrito acima); o algoritmo **PF** (Propagation/Filtration), similar ao DRed; o algoritmo **Kuchenhoff**, o qual gera regras para executar um cálculo da diferença entre estados consecutivos da base de dados e o algoritmo Urpi-Olive, que executa regras de transição refletindo as modificações em cada relação derivada de cada alteração na relação base.

5.3 SOLUÇÃO PROPOSTA

A partir do estudo desses dois algoritmos, foi definido o uso do algoritmo de ***counting***, devido possuir mais flexibilidade, já que pode ser utilizado em visões cujas mesmas possuem ou não linhas duplicadas; e por utilizarmos visões não recursivas, as quais como visto anteriormente, tem como técnica mais apropriada o algoritmo de ***counting***.

Uma vez rodado o algoritmo escolhido percebemos a duplicidade de tuplas geradas pelo mesmo, já que o algoritmo de ***counting*** permite a execução de linhas duplicadas. Isto pode ser notado a partir do conteúdo das variáveis contadoras.

A seguir temos o trecho de código responsável pela criação da visão e pela contagem regressiva de cada tupla derivada e duplicada da visão.

```
// CRIAÇÃO DA VIEW SEER
$query="CREATE VIEW seer AS SELECT
        nArticleID,chMetaTitle,chMetaSubject,chMetaAbstract,dtDatePublished,fkIssueID FROM
        tblarticles";
$result=mysql_query($query,$link);
$query2="SELECT nArticleID,chMetaTitle,chMetaSubject,chMetaAbstract,dtDatePublished,fkIssueID
        FROM tblarticles";
$result2=mysql_query($query2,$link);
if(mysql_num_rows($result)>0)
{
  $aux=0;
  while($campos=mysql_fetch_array($result))
  {
    // CRIAÇÃO DA VARIÁVEL COUNTING
    $counting[$aux]=0;
    while($campos2=mysql_fetch_array($result2))
    {
      if($campos['chMetaTitle']==$campos2['chMetaTitle'])
        $counting[$aux]++;
    }
    $aux++;
```

Como vemos a variável $counting[$aux] armazena para cada registro verificado a quantidade de vezes que o mesmo foi duplicado. Para isso verificamos as variáveis $campos['chMetaTitle'] e $campos2['chMetaTitle'], que representam os títulos dos trabalhos contidos nos sistemas; pois na função ***while*** fazemos uma varredura completa para sabermos quais títulos de trabalhos são iguais, ou seja, para cada titulo de trabalho executamos uma varredura por todos os outros títulos, em busca de algum outro título igual. Uma vez executada esta pesquisa a variável $counting[$aux] armazena a quantidade de derivações iguais para cada titulo de trabalho. A partir do

conteúdo da variável $counting[$aux] sabemos a duplicidade de cada registro da visão gerada.

Dando continuidade à execução do algoritmo, uma derivação de cada uma das tuplas geradas precisa ser excluída. Isto pode ser observado no trecho de código abaixo, onde excluímos uma derivação duplicada de cada registro gerado de acordo com o conteúdo dos contadores.

```
// REMOÇÃO DE TUPLAS DERIVADAS

for($i=0;$campos=mysql_fetch_array($result);$i++)
{
   if($counting[$i]>0)
   {
      $query="DELETE FROM seer WHERE chMetaTitle=".$campos['chMetaTitle']."AND
                  nArticleID=".$campos['nArticleID'];
      $result2=mysql_query($query,$link);
   }
}
```

Como vemos, se a variável $counting[$i] for maior que zero, então sabemos que aquele registro identificado pelo índice $i tem mais de uma derivação, logo será necessário excluir uma duplicidade do registro aqui citado.

É importante ressaltarmos que devemos apagar todas as derivações de cada tupla, antes de finalizar o algoritmo, já que não podemos ter registros duplicados no resultado final da visão.

Vale informa que o algoritmo deve ser rodado toda vez que a relação base for alterada, dessa forma a visão gerada estará sempre em conformidade com os dados oriundos da base de dados.

O mesmo algoritmo foi executado para criação da visão materializada das teses e dissertações do VIRTUA, formando assim o escopo do DSPACE.

Dessa forma podemos concluir, que com o uso do algoritmo de ***counting*** as visões são geradas sem replicação de dados, construindo um modelo formal de integração de sistemas de informação. Assim na solução atual, podemos observar o modelo de integração proposto como alternativa para os sistemas sempre

permanecerem interligados, através da execução contínua do algoritmo, garantindo desse modo a fidelização precisa dos dados gerados pelas visões em conformidade com a relação base.

Na figura 4 podemos observar o novo modelo proposto.

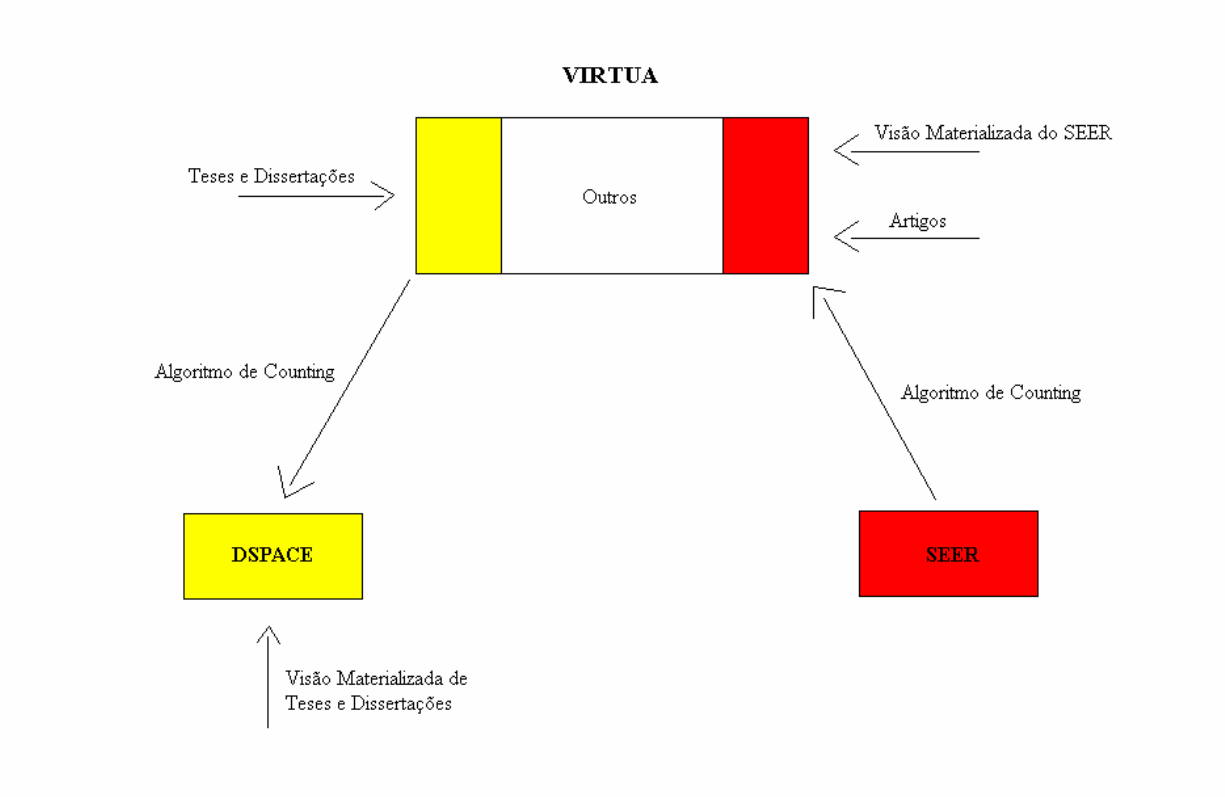

Figura 4 – Estrutura de Integração Proposta

Como podemos ver o SEER executa o algoritmo de ***counting***, criando assim uma visão materializada do SEER no VIRTUA, em seguida o VIRTUA executa o algoritmo, construindo uma visão materializada das teses e dissertações encontrados no VIRTUA para o DSPACE. Dessa forma padronizamos todo o envio de metadados através da execução do algoritmo, garantindo assim a fidelidade das informações extraídas das relações base, ou seja, as visões geradas sempre estarão em conformidade com as informações armazenadas na base de dados original. Assim podemos afirmar que o algoritmo proporciona a exatidão dos dados replicados.

Abaixo temos a consulta responsável pela ligação SEER - VIRTUA.

create view seer as select nArticleID,chMetaTitle,chMetaSubject,chMetaAbstract,dtDatePublished, fkIssueID form tblarticles;

A partir desta consulta os dados do SEER são copiados para o VIRTUA, construindo assim uma visão materializada do SEER.

Abaixo temos a consulta responsável pela ligação VIRTUA – DSPACE

create view virtua as select distinct f.bib_id from bibliographic_search s, bibliographic_fields f ,BIB_SUBJECT_V b where contains(s.search_keywords,'\@T within ty') > 0 and s.bib_id = f.bib_id and f.bib_id=b.bib_id and b.subject = 'TXCD' and f.creation_date >= to_date('18082004','ddmmyyyy') and (f.modify_date >= trunc(sysdate) - interval '2' day or f.creation_date >= trunc(sysdate) - interval '2' day);

Com isso nós temos uma replicação dos metadados referentes a teses e dissertações encontrados no VIRTUA, ou seja, a partir desta consulta é construída uma visão materializada de parte das informações encontradas no VIRTUA.

Para fins de comprovação, na tabela 2 listaremos alguns dados usados antes da aplicação do algoritmo. É importante citar que vários registros foram adicionados exclusivamente com o intuito de realizar testes, ou seja, após a conclusão do trabalho os mesmos foram excluídos pelo próprio algoritmo.

nArticleID	chMetaTitle
100	Submissão sem Título
101	Submissão sem Título
104	CARACTERIZAÇÃO QUÍMICA DO COGUMELO AGARICUS BLASEI MURRIL
105	LIBERAÇÃO DE FERRO (III) DE MICROESFERAS RETICULADAS DE QUITOSANA
107	CLASSIFICAÇÃO DE PRODUTOS DE ORIGEM VEGETAL
112	Estudo Alelopático Aplicado de Aster lanceolatus, Willd.
115	FOSFINA: RISCOS
118	Apenas um teste
119	Apenas um teste

Tabela 2 – Dados Pré-Utilizados.

Como vemos, a duplicidade de dados é clara, como exibi as duas primeiras e as duas últimas linhas da tabela.

Abaixo na tabela 3 temos o resultado dos dados depois da aplicação do algoritmo.

nArticleID	chMetaTitle
100	Submissão sem Título
104	CARACTERIZAÇÃO QUÍMICA DO COGUMELO AGARICUS BLASEI MURRIL
105	LIBERAÇÃO DE FERRO (III) DE MICROESFERAS RETICULADAS DE QUITOSANA
107	CLASSIFICAÇÃO DE PRODUTOS DE ORIGEM VEGETAL
112	Estudo Alelopático Aplicado de Aster lanceolatus, Willd.
115	FOSFINA: RISCOS
118	Apenas um teste

Tabela 3 – Dados Pós-Utilizados.

Na tabela 3 podemos observar claramente a falta de duplicidade dos registros, ou seja, com a aplicação do algoritmo de ***counting*** as tuplas com nArticleID igual a 101 e 119 foram excluídas, garantindo assim, a não replicação de dados e a contínua fidelização da informação em conformidade com a relação base.

Podemos usar a tabela 2 para dar um exemplo mais genérico. Se nós atribuirmos letras para cada título da tabela 2, teremos a seguinte relação *link*: {(100,a), (101,a), (104,b), (105,c), (107,d), (112,e), (115,f), (118,g), (119,g)}. Se ignorarmos a diferença das chaves, teremos link = {(T,a), (T,a), (T,b), (T,c), (T,d), (T,e), (T,f), (T,g), (T,g)}. Como resultado desta relação, temos a visão hop = {(T,a), (T,b), (T,c), (T,d), (T,e), (T,f), (T,g)}. As tuplas hop (T,a) e hop (T,g) possuem duas ocorrências, pois são derivadas duas vezes de link. Como não estamos usando o operador ***distinct***, ou seja, a relação link possui duplicidade, então teremos ***count*** igual a 2 para a tupla hop (T,a) e ***count*** igual a 2 para a tupla hop (T,g). De acordo com o algoritmo de ***counting*** uma derivação de cada uma das tuplas hop (T,a) e hop (T,g) precisa ser excluída. Dessa forma podemos afirmar que o algoritmo de ***counting*** funciona para qualquer tipo de dado, desde que o mesmo esteja estabelecido em uma relação base.

Como benefícios dessa nova modelagem podemos citar: a falta de duplicidade das informações geradas pelas visões materializadas, facilitando assim o acesso direto aos dados requeridos; a constante garantia de fidelidade e integridade das informações em relação aos dados base, permitindo dessa maneira o usuário ter certeza que a informação que o mesmo acessa está sempre atualizada de acordo com a base de dados original; o uso de visões materializadas como uma cópia de segurança, dessa forma os dados estarão sempre duplicados garantindo assim a contínua existência das informações armazenadas na relação base e nas visões materializadas; a utilização de um algoritmo consolidado de visões materializadas como prova matemática na modelagem formal de um projeto de integração de sistemas de informação heterogêneos, viabilizando a construção de um padrão de integração de diferentes sistemas de informação, através do uso de visões materializadas.

6 CONCLUSÃO

Atualmente as Bibliotecas Digitais e demais Sistemas de Informação são de suma importância para a divulgação da informação, ultrapassando obstáculos, como os altos custos financeiros impostos pelas entidades privadas e as dificuldades na disseminação de conteúdo. Para tais vantagens é necessário que os Sistemas de Informações, como sistemas de automatização do acervo físico das bibliotecas convencionais, Bibliotecas Digitais, Revistas Eletrônicas e demais estruturas, utilizem mecanismos semelhantes na sua forma de manuseamento da informação.

O protocolo OAI-PMH esta cada vez mais se consolidando como protocolo padrão para bibliotecas digitais e repositórios de dados científicos. Isso pode ser visto através do constante surgimento de projetos que adotam o OAI-PMH, como padrão a ser usado, e no desenvolvimento de interfaces para base de dados e bibliotecas já existentes.

Como foi possível observar através da estrutura de integração modelada, percebemos a grande importância do conceito de visões materializadas, uma vez que a definição das mesmas foi empregado nos sistemas de informação da UFPR para construir toda uma infra-estrutura de integração entre os diferentes sistemas.

O objetivo inicialmente proposto, de modelar uma estrutura de integração de sistemas de informação fazendo uso do conceito de visões materializadas foi atingido e comprovado através dos resultados exibidos pelos diferentes sistemas de informação situados na UFPR. A partir desta iniciativa foi possível identificarmos os pontos essenciais para a execução de um modelo de integração de sistemas de informação e propor a divulgação do mesmo para instituições e centros de ensino e pesquisa que sofrem com a falta de integração de seus meios de armazenamentos digitais.

Este trabalho provê as instituições brasileiras e de outros países mais um recurso de pesquisa centralizada, incentivando assim o acesso à produção científica e a centralização das informações, através da integração de dados de sistemas situados em diferentes localidades remotas do globo.

7 TRABALHOS FUTUROS

A partir deste trabalho podemos desencadear diversos outros projetos de pesquisa, uma vez que este modelo abrange apenas uma parte dos setores físicos da UFPR. A idéia seria expandir esta estrutura de integração para alcançar demais setores da universidade, criando assim blocos de informação integrados, espalhados por todos os departamentos da UFPR.

Como futuras pesquisas, devemos citar também a possibilidade de ser criado um padrão de integração de informação em cima de visões materializadas. Tal padrão seria de suma importância para universidades, centros de pesquisa e órgãos que desejem a centralização de seus dados em uma estrutura integrada. Dessa forma a criação desse padrão geraria a possibilidade de implantarmos um mecanismo de busca unificada, onde tal busca percorreria escolas, centros de pesquisa, empresas e demais órgãos que estivessem com suas bases de dados em conformidade com o padrão aqui proposto.

É importante ressaltarmos a importância da divulgação deste trabalho, para que demais órgãos possam implantar projetos semelhantes a fim de terem seus núcleos de informação unidos em uma só estrutura. A ampla divulgação é bastante necessária para o esclarecimento de dúvidas e para possibilidade de refinamento na execução desse mecanismo de integração. Assim todos que tiverem acesso às informações sobre este trabalho, poderão optar sobre o mesmo, retirando *bugs* e adicionando novas idéias, cujas mesmas irão ajudar para melhorar cada vez mais a execução e finalidade deste projeto.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] CATHRO, W.S. **Digital Libraries: A National Library Perspective**. Disponível em: <http://www.nla.gov.au/nla/staffpaper/cathro4.html> Acesso em: 13 de setembro de 2005.

[2] TAYLOR, C. **Cibrary An Introduction to Metadata**. Disponível em: <http://www.library.uq.edu.au/iad/ctmeta4.html> Acesso em: 15 de setembro de 2005.

[3] VAN, H.S.; YOUNG, J. A.; HICKEY, T. B. **Using the OAI-PMH … Differently**. Disponível em: <http://www.dlib.org/dlib/july03/young/07young.html> Acesso em: 20 de setembro de 2005.

[4] LAGOSE, C.; VAN DE SOMPEL, H. **The Open Archives Initiative: Building a low-barrier interoperability framework**. Disponível em: <http://www.cs.cornell.edu/lagoze/papers/oai-jcdl.pdf> Acesso em: 22 de setembro de 2005.

[5] CRISTINA, I. I.; EDUARDO, J. F. **O Uso de Visões Materializadas em Data Warehouses**. Disponível em: <http://www.ime.usp.br/~am/5701/isabel-relat.pdf> Acesso em: 3 de fevereiro de 2006.

[6] WILSON, C. D. A; FALCÃO, D. C. L; FABRÍCIO, F. M; DIAS, G. A; BOSCO, J. D. J; COLASCO, J. M. A. A. **Desenvolvendo uma ferramenta voltada para a gerência do processo de publicação de periódicos científicos eletrônicos na World Wide Web através do uso de soluções baseadas em software livre**. 55ª Reunião Anual da SBPC – Recife – PE, 2003.

[7] HARTER, S. P. **What is a Digital Library? Definitions, Content, and Issues**. Disponível em: <http://php.indiana.edu/%7Eharter/korea-paper.htm> Acesso em: 8 de outubro de 2005.

[8] COPYLEFT. **Entenda o que é o conceito "copyleft"**. Disponível em: <http://www1.folha.uol.com.br/folha/informatica/ult124u12307.shtml> Acesso em: 1 de outubro de 2005.

[9] RODRIGUES, C. L. **O Conceito de Bibliotecas nas Bibliotecas Digitais The Concept of Library in Digital Libraries**. Disponível em: <http://www.informacaoesociedade.ufpb.br/html/IS1420401/> Acesso em: 17 de outubro de 2005.

[10] HARNAD, S. **Nature Debates: The Self-Archiving Initiative**. Disponível em: <http://www.nature.com/nature/debates/eaccess/Articles/harnad.html> Acesso em: 15 de outubro de 2005.

[11] OAI-Tech. **Open Archives Initiative – Technical Committee.** Disponível em: <http://www.openarchives.org/organization/index.html> Acesso em: 19 de setembro de 2005.

[12] DCMI. **The Dublin Core Metadata Initiative**. Disponível em: <http://www.dublincore.org/> Acesso em: 21 de setembro de 2005.

[13] DCES. **Dublin Core Metadata Element Set, V 1.1: Reference Description**. Disponível em: <http://dublincore.org/documents/dces/> Acesso em: 19 de outubro de 2005.

[14] OAI. **Open Archives Initiative**. Disponível em: <http://www.openarchives.org/> Acesso em: 10 de outubro de 2005.

[15] OAIster. **University of Michigan Digital Library Production Service**. Disponível em: <http://oaister.umdl.umich.edu/o/oaister/> Acesso em: 19 de setembro de 2005.

[16] W3C. **World Wide Web Consortium**. Disponível em: <http://www.w3.org/> Acesso em: 26 de setembro de 2005.

[17] PASQUAL, J. **Uso de XML para Interoperabilidade entre Bases Heterogêneas**. Dissertação de Mestrado. Curitiba, 2002. Setor de Ciências Exatas. Universidade Federal do Paraná.

[18] PINTO, J. S. P. **Mapeamento de Atributos Complexos e Multivalorados na Extração de Esquemas utilizando XML**. Dissertação de Mestrado. Curitiba, 2001. Setor de Ciências Exatas. Universidade Federal do Paraná.

[19] LAGOSE, C. **The Open Archives Initiative Protocol for Metadata Harvesting**. Disponível em: <http://www.openarchives.org/OAI/openarchivesprotocol.html> Acesso em: 18 de outubro de 2005.

[20] LAGOZE, C.; VAN DE SOMPEL, H. **Notes from the Interoperability Front**. Disponível em: <http://www.openarchives.org/documents/ecdl-oai.pdf> Acesso em: 27 de outubro de 2005.

[21] VTLS. **Virginia Tech Library System**. Disponível em: <http://www.vtls.com/> Acesso em: 23 de setembro de 2005.

[22] DSPACE. **The DSpace Digital Repository**. Disponível em: <http://dspace.org/index.html> Acesso em: 16 de setembro de 2005.

[23] SEER. **Sistema Eletrônico de Editoração de Revistas**. Disponível em: <http://www.ibict.br/secao.php?cat=SEER> Acesso em: 29 de setembro de 2005.

[24] GIACOMO, A. P.; CAROLINO, G. S.; APARECIDA, M. D. P.; APARECIDA, C. R. **Implementação de um Módulo de Circulação Através do Software de Funções Integradas VIRTUA/VTLS: a experiência do Sistema de Bibliotecas da Unicamp**. Disponível em: <http://libdigi.unicamp.br/document/?code=7> Acesso em: 28 de outubro de 2005.

[25] KENZIE, M. S.; BASS, M.; MCCLELLAN, G.; TANSLEY, R.; BARTON, M.; BRANSCHOFSKY, M.; STUVE, D.; HARFORD, J. W. **DSpace An Open Source Dynamic Digital Repository**. Disponível em: <http://www.dlib.org/dlib/january03/smith/01smith.html> Acesso em: 12 de outubro de 2005.

[26] BRITO, S. C. S.; FERREIRA, S. A.; ÁNGEL, M. M. A.; CAROLINO, G. S. **I Workshop Virtual Cibereduc – Seer: Periódicos Eletrônicos: Editoração e Acesso**. Disponível em: <http://eprints.rclis.org/archive/00003347/01/ETD-2005-30.pdf> Acesso em: 13 de outubro de 2005.

[27] SACCOL, D. B.; HEUSER, C. A. **Materialização de Visões XML**. Disponível em: <http://www.inf.ufrgs.br/~deise/clei.pdf> Acesso em: 21 de outubro de 2005.

[28] RONCONI, V. **O Otimizador do Oracle para Desenvolvedores Parte II – Otimizador Baseado em Custo**. Disponível em: <http://www.sqlmagazine.com.br/artigos/oracle/08_OtimizadorOracle-II.asp> Acesso em: 2 de outubro de 2005.

[29] CRAVEIRO, J. C. N.; CRESTANA, H. G.; MATSUMOTO, L. S. **Integração de um Banco de Dados e um Data Warehouse sobre um Sistema de Arquivos Paralelos**. Disponível em: <http://www.dct.ufms.br/~craveiro/public/wscad01ibddwsap.pdf> Acesso em: 29 de setembro de 2005.

[30] ALBERTO, C. A. M.; MANOEL, L. S. C. **Implantação do Armazém de Dados da Fruticultura no Ministério da Agricultura, Pecuária e Abastecimento**. Disponível em: <http://www.cnptia.embrapa.br/modules/tinycontent3/content/2001/INSTRTECNICAS8int.pdf> Acesso em: 10 de setembro de 2005.

[31] LIU, X.; MALY, K.; ZUBAIR, M.; NELSON, M. L. **Repository Synchronization in the OAI Framework**. Disponível em: <http://whiskey.cs.odu.edu/~liu_x/paper/freshness/freshness.pdf> Acesso em: 12 de maio de 2006.

[32] PI. **Portal da Informação da UFPR**. Disponível em: <http://www.portal.ufpr.br/> Acesso em: 31 de julho de 2006.