METADADOS PARA A DESCRIÇÃO DE RECURSOS DA INTERNET: As novas tecnologias desenvolvidas para o padrão *Dublin Core* e sua utilização

Ana Maria Pereira;

Divino Ignácio Ribeiro Júnior;

Guilherme Luiz Cintra Neves

RESUMO

A presente pesquisa teve como objetivo desenvolver o conhecimento necessário à produção de software quanto à recuperação da informação através das tecnologias disponíveis para a Internet (XML, JAVA ou PHP, RDF, Dublin Core), verificar o funcionamento do padrão de metadados Dublin Core e a estrutura de descritores RDF (Resource Description Framework), que serão a base de sustentação no desenvolvimento desse software. Identificar as linguagens que serão usadas na criação do software (XML, JAVA ou PHP), para subsidiar o desenvolvimento do mesmo, além de analisar a sintaxe e o uso das tecnologias disponíveis, na aplicação do padrão Dublin Core, para a recuperação da informação, por meio dos Metatags do HTML, tecnologia XML e da estrutura de descrição de recursos RDF. A pesquisa foi do tipo levantamento bibliográfico, e avaliou a disponibilidade dessas ferramentas para o usuário comum na rede. Como considerações finais, verificou-se que as análises e estudos dos metadados e das tecnologias disponíveis permitiram descrever como construir uma interface de busca baseada no padrão Dublin Core e na tecnologia XML, contribuindo assim, para aplicações práticas e benefícios que o estudo de Metadados proporciona tanto para o meio acadêmico, quanto para a sociedade.

Palavras-chave: Recuperação da Informação; Metadados; *Dublin Core*; Mecanismos de Busca

1 INTRODUÇÃO

Este artigo apresenta a utilização do padrão *Dublin Core* (DC) para a Recuperação da Informação na Internet, bem como as tecnologias disponíveis para sua aplicação (*eXtended Markup Language* (XML), *Resource Description Framework* (RDF), *Definition Type Document* (DTD), entre outras).

Pretende ainda, demonstrar a aplicação dessas tecnologias e sua utilização, a integração dos elementos fundamentais para a Recuperação da Informação na Internet e propor um modelo de software que permita recuperar documentos que aplicam o padrão DC como recurso descritivo.

O objetivo principal dessa pesquisa é o de subsidiar a implementação de recursos de software de forma a dispor de um mecanismo de busca capaz de localizar e manter uma base de referências de documentos que aplicam o padrão DC como recurso descritivo.

Para tanto, foram realizados estudos sobre o recurso descritivo *Resource Description Framework*, a sintaxe e o uso da tecnologia XML na aplicação do padrão DC para representação descritiva e qualitativa de recursos da Internet, e da aplicação deste padrão diretamente as *metatags* do código HTML de documentos da Internet. Também foram identificados s*ites* que implementam o padrão DC, a fim de analisar os recursos de interface de busca comumente aplicados.

Ao abordar a revisão de literatura sobre metadados e as tecnologias estudadas, foi possível observar que os padrões definidos para metadados têm como características a recuperação do item documentário. Uma das motivações desta pesquisa é a de buscar subsídios para, não só recuperar o item documentário, mas, também, recuperar seu conteúdo,

contribuindo, assim, com conhecimento diretamente aplicável no contexto profissional e de ensino da área de Biblioteconomia.

Uma das preocupações da área de Ciência da Informação é facilitar o acesso à informação, como forma de recuperação e disseminação da informação, e é nesse sentido que o uso de tecnologias existentes são periodicamente aplicadas, no intuito de aprimorar seus benefícios para a área.

Os estudos relacionados aos processos de automação proporcionaram uma melhor interação entre bibliotecas nacionais e internacionais, concretizando, assim, o intercâmbio informacional.

Um sistema de intercâmbio é capaz de suprir deficiências e otimizar a aplicação de recursos, e é com base nesse pressuposto que se verifica o compartilhamento de informação como algo inovador, sobretudo para se estabelecer padrões ao enorme volume de documentos eletrônicos vigentes.

Com esse propósito, foi realizado em Ohio, EUA, em 1995, o *Dublin Metadata Wokshop*. O resultado desse workshop representa um recurso simples de descrição de arquivo, com potencial para melhorar o acesso à informação na internet. (FROST, 2002).

Após o levantamento bibliográfico decidiu-se utilizar para experimentação uma distribuição do sistema operacional *Linux*, com o *software* livre *HtDig* e um servidor web *Apache*, para disponibilizar uma interface de busca que recupere informações a partir do padrão DC. Com o uso de uma ferramenta de busca já existente e amplamente utilizada como o *HtDig*, tornou-se desnecessária a utilização da linguagem de programação PHP, inicialmente prevista no projeto como linguagem para a construção dessa interface.

Por meio de pesquisas na Internet identificou-se diversas ferramentas para a recuperação de informações na Internet, várias delas aplicando linguagem PHP, entretanto, desde a fase de projeto desta pesquisa, propôs-se o uso da ferramenta *HtDig* (programada em linguagem C, e disponível somente para plataformas *Linux*), que oferece todas as funcionalidades necessárias para trabalhar com meta-tags com descritores DC.

As vantagens oferecidas pelos desenvolvedores do *HtDig*, que culminaram na decisão de utilizá-lo como base para a realização de teste, são a ampla documentação, disponível no sítio *web* http://www.htdig.org, a política de código fonte aberto, a política de licença pública (GPL/GNU) e ser desenvolvido para ambiente *Linux*, o que garante uma estrutura sólida e confiável para o *software.*

2 REVISÃO DE LITERATURA

Os metadados constituem-se importantes ferramentas para a descrição do conteúdo de um determinado conjunto de dados de um item informacional, em rede eletrônica. A padronização possibilita o fácil acesso e recuperação da informação e os usuários podem mover com facilidade os dados entre os diferentes sistemas e plataformas informáticas.

2.1 Conceito de Metadados

Ao estudar os metadados e sua criação utilizando a linguagem XML, é possível analisar sua operabilidade para todo e qualquer tipo de informação disponível na WEB. Preocupou-se também, com a recuperação

da informação no ambiente telemático, visto que a disseminação da informação depende do processo de armazenamento, padronização e recuperação.

Mesmo sendo atual somente a partir da década de 90, o termo *Metadata*, não é tão recente. A palavra *Metadata* foi criada por Jack Myres em 1969, para denominar os dados que descreviam registros de arquivos convencionais.

Para Alvarenga (2003, p. 19), o conceito de metadado,

> etimologicamente, quer dizer '*dado sobre dado*' ; dado que descreve, a essência, atributos e contexto de emergência de um recurso (documento, fonte, etc.) e caracteriza suas relações, visando-se ao acesso e ao uso potencial. O prefixo grego *meta significa mudança, posterioridade, além, trancendência* [...].

Duval et al. (2002, p. 11, tradução nossa), define metadados como sendo dados *estruturados* sobre dados. Metadados ou *Metadata* é a descrição de dados sobre dados, ou seja, segundo Sumpter (1994 apud GOMES, 2001, p.5), "é a informação sobre o dado que permite o acesso e gerenciamento deste dado de maneira eficiente e inteligente".

De acordo com Garcia (1999, p. 3),

> podem ser destacadas as seguintes vantagens na utilização e disponibilização de metadados:
>
> a) estabelecimento de padrões de dados diante da heterogeneidade de informações contidas na rede, como por exemplo a Internet;
> b) facilidade e maior precisão na recuperação das informações desejadas, troca de informações entre aplicações e organizações.

Os metadados incluem elementos de descrição do conteúdo dos dados e qualquer informação relevante para a recuperação informacional dos mesmos. Os metadados tem como vantagens, segundo Souza, (1997):

a) possibilitar a interoperabilidade entre as diversas fontes de dados;

b) definir a linguagem de consulta;

c) permitir a agilidade e o acesso com qualidade na recuperação da informação;

d) e, propiciar o intercâmbio informacional.

De acordo com Pereira (2001), entre o uso dos metadados, podemos distinguir os mais significativos:

a) manter o investimento na organização interna dos dados geoespaciais;
b) providenciar informação sobre dados existentes sobre determinada área de interesse, localização desses dados, grau de atualização dos dados, formato e obstáculos à sua utilização;
c) e, providenciar informação necessária para processar e interpretar dados recebidos de uma fonte exterior.

Após verificar as vantagens e o uso dos metadados em ambientes de rede (intranet ou Internet), foi possível compreender sua importância para a Recuperação da Informação na *Web*, e sua contribuição para a sociedade.

Ainda, segundo Pereira (2001, p. 2), os metadados desempenham quatro funções:

a) acessibilidade – dados necessários para determinar os conjuntos de dados existentes para uma determinada localização geográfica;
b) compatibilidade de uso – dados necessários para determinar se um conjunto de dados se enquadra em determinado fim;
c) acesso – dados necessários para que se adquira um conjunto de dados identificados;
d) transferência – dados necessários para processar e usar um conjunto de dados.

Dadas as funções que os metadados apresentam, é preciso entender sua utilização no contexto da Internet, onde são desenvolvidos e utilizados os mais diferentes padrões de metadados, criando assim a necessidade de integrá-los. Com os diferentes padrões de metadados existentes, e com o objetivo de sanar as necessidades de seus usuários, de acordo com as finalidades distintas de informações às quais estejam associados, foram desenvolvidos vários modelos . Entre eles, pode-se citar:

a) FGDC – *Federal Geographic Data Committee* – para descrição de dados geoespaciais;
b) MARC – *Machine Readable Catalogue* – para catalogação bibliográfica;
c) *IAFA/WHOIS++ - Internet Anonymous Ftp Archive with transfer protocol* – para descrição do conteúdo e serviços disponíveis em arquivos *ftp – file transfer protocol;*
d) *TEI – Text Encoding Initiative* – para representação de materiais textuais na forma eletrônica;
e) DC – DC – para catalogação de documentos eletrônicos na *Web;*

f) *SAIF – Spatial Archive and Interchange Format* – para compartilhamento de dados espaciais e espaço-temporais.

2.2 O Padrão Dublin Core

No contexto atual de produção, organização e recuperação da informação em ambiente Web, as metas de trabalho não podem se restringir à criação de representações simbólicas dos itens bibliográficos em suportes físicos, constantes de um determinado acervo, mas compreendem estabelecimento dos denominados metadados, muitos do quais podem ser extraídos diretamente dos próprios objetos, constituindo-se os mesmos como chaves de acesso a serviço de todo e qualquer usuário da informação no espaço cibernético. Com esse propósito, foi realizado na cidade de Ohio, USA, em 1995, o *I Dublin Metadata Workshop*.

O resultado desse *Workshop* representa um recurso simples de descrição de arquivo, com potencial para otimizar o acesso à informação na Internet. A principal finalidade dos metadados é documentar e organizar de forma estruturada os dados nas Unidades de Informação (UI), com o objetivo de minimizar a duplicação de esforços e facilitar a manutenção dos dados.

A tecnologia de metadados está surgindo em função das necessidades das UIs de conhecer melhor os dados que mantêm e conhecer com mais detalhes os dados de outras UIs. Os dados precisam conter informações que auxiliem os usuários a tomar decisões sobre a sua devida aplicação. A catalogação com essa base permitirá a maior utilização destes dados por usuários com múltiplos interesses.

O *Dublin Core Metadata Initiative* (DCMI), é um projeto destinado a organizar as informações nas páginas da WEB, com o objetivo de estabelecer padrões de catalogação e classificação das informações no meio eletrônico. O DC, tem suas origens em Chicago, na 2ª. Conferência Internacional sobre a WWW em outubro de 1994, quando Yuri Rubinsky, Stuart Weibel e Eric Miller integrantes da OCLC – *Online Computer Library Center* e Joe Hardin da NCSA – *National Center for Supercomputing Applications*, iniciaram uma discussão em semântica e WEB.

Essa iniciativa levou a NCSA e a OCLC a organizarem em 1995 um evento denominado de OCLC/NCSA Metadata Workshop, onde os participantes discutiram um conjunto semântico para recursos baseados na WEB, com o propósito de agilizar a pesquisa e recuperação de recursos informacionais na WEB. O objetivo principal desse *workshop* era definir um conjunto mínimo de elementos de descrição para recursos da WEB.

Segundo Desai (1997), pretendia-se tratar o problema da catalogação de recursos da rede, com a adoção, a extensão ou a modificação de padrões existentes e de protocolos para facilitar a recuperação e o acesso a informação, utilizando os elementos de metadados.

O DC apresenta um conjunto de 15 elementos de metadados, considerados mínimos para facilitar a recuperação da informação do documento eletrônico. Segundo Souza (2000, p. 94), são eles: "título; criador; assunto; descrição; publicador; colaborador; data; tipo; formato; identificador; fonte; idioma; relação; cobertura; direito autoral". Esses elementos são considerados mínimos para a representação de um item informacional em ambiente telemático.

Com a expansão da Internet e conseqüente desenvolvimento da tecnologia de redes eletrônicas, o volume de informações cresce

desordenadamente. Portanto, torna-se imprescindível o desenvolvimento de padrões que visem a descrição dos recursos de informação. No domínio das bibliotecas tradicionais, temos a ISO 2709 que assume, como formato, o intercâmbio universal para registros bibliográficos. No contexto das bibliotecas digitais surge o *Dublin Core Metadata Initiative*, com o propósito de discutir e propor padrões de descrição de recursos informacionais.

Segundo Souza (2000, p. 93),

> o *Dublin Core*, pode ser definido como sendo o conjunto de elementos de metadados planejado para facilitar a descrição de recursos eletrônicos. A expectativa é de que os autores e Websites, que não possuam conhecimentos em catalogação, possuam capacidade de usar o *Dublin Core* para descrição de recursos eletrônicos, tornando suas produções mais visíveis aos mecanismos de busca e sistemas de recuperação.

Para atingir tal objetivo é preciso divulgar o padrão DC por meio de uma estratégia de marketing consistente, orientada ao usuário (autores, webmasters, instituições), que disponibiliza informações na Internet.

Para Baptista (2001, p. 78), "o *Dublin Core* é um conjunto de metadados cujo objetivo é facilitar a descoberta de recursos eletrônicos", e suas características são: "simplicidade, interoperabilidade semântica, consenso internacional, extensibilidade e modularidade de metadados na Web".

Com a padronização de elementos mínimos para descrição de recursos online e sua ampla implementação nos sítios da Internet,

certamente as buscas retornarão resultados com baixa revocação e alta relevância.

Barreto (apud ALVES, 2002, p. 79) afirma, sobre o padrão DC, que "sua simplicidade é o fator chave para a rápida utilização do padrão [...] na forma de uma aplicação integrando todos os tipos de informação, inclusive aqueles não disponíveis no meio eletrônico".

Enquanto a interoperabilidade semântica permite a comunicação com diferentes padrões, o consenso internacional é um indicativo da ampla utilização do padrão DC pela comunidade científica, que por ser flexível, permite a extensibilidade e modularidade dos metadados.

Sendo o DC, de acordo com Alves (2002, p. 36), uma "[...] iniciativa entre profissionais de diversas áreas no intuito de resolver o problema de descrição dos recursos na rede [...]", além de ser uma recomendação do W3C – "[...] o DCMI possui uma forte ligação com o W3C e com comunidades de desenvolvedores em RDF e XML, possibilitando a essas ferramentas possuir codificação para o padrão DC [...]" (GRÁCIO, 2002, p. 39) –, e das definições acompanhadas anteriormente, podemos inferir as vantagens advindas do uso desse padrão na Recuperação de Informações na Internet.

Citando Duval et al. (2002, tradução nossa), podemos dizer que existem diferentes possibilidades de se implementar metadados, seja por parte dos criadores do recurso (descrição simples e sujeita a erros no processo de descrição do recurso), por profissionais da informação (descrição pode ser mais complexa e fidedigna, porém aumenta o custo de implantação dos metadados), e por aplicações que podem facilitar a descrição (tornando evidente que certos campos não precisam ser utilizados), aplicações que auxiliem o usuário a selecionar os valores

apropriados para um elemento em particular, e finalmente, aplicações que podem identificar valores para alguns elementos com maior confiabilidade do que o usuário. A riqueza de descrição do metadado será determinada por políticas e práticas determinados pela agencia criadora do metadado, e serão guiadas pelos requisitos funcionais de serviços ou aplicações.

Ainda, Duval et al. (2002, p. 11, tradução nossa) apresenta também as características sobre as descrições de metadados detalhadas: podem aumentar a precisão da busca; requerem alto investimento na criação do metadado; torna mais difícil promover a consistência na criação de metadados. Mas, por outro lado, apresenta as características das descrições simples: são mais fáceis e baratas de serem geradas; podem apresentar mais resultados incorretos ou inconsistentes, ou maior esforço dos usuários para identificar os resultados mais relevantes; melhora a probabilidade da interoperabilidade entre disciplinas.

O conjunto de descritores do DC pode estar intrínseco no próprio documento descrito – por meio da linguagem HTML (*Hiper Text Markup Language*), XML e outras, ou, de acordo com o recurso necessário, a meta-informação pode estar separada do recurso utilizado para a catalogação.

Duval et al. (2002, p. 9, tradução nossa) esclarece as diferenças entre essas duas formas de descrição dos recursos, o *embedded metadata* (metadado intrínseco no recurso) e o *associated metadata* (metadado extrínseco ao recurso):

a) *embedded metadata* pode ser *garimpado* (ou seja, pode ser rastreado na internet por um motor de busca), o que aumenta sua visibilidade e pode incentivar os criadores a implementar metadados em seu recurso;

b) *associated metadata* é mantido em arquivos associados ao recurso que descrevem, eles podem ou não serem *garimpados*, a principal vantagem desse tipo de metadado é que ele facilita o controle sem alterar o conteúdo do recurso, mas esse benefício é pago ao custo da simplicidade, exigindo co-gerenciamento dos arquivos de recursos e dos arquivos de metadados.

c) Nesta categoria, existe ainda o *third-part metadata* que é mantido em um repositório separado do recurso, por uma organização que pode ou não ter controle direto ou acesso ao conteúdo dos recursos, geralmente esse tipo de metadado é mantido em um banco de dados inacessível aos motores de busca.

O primeiro *workshop*, realizado em Dublin, Ohio, teve três objetivos principais:

a) identificar a clientela ou comunidades envolvidas com recuperação de documentos;

b) imaginar as descrições de metadados em potencial que pudessem atender essa população de usuários, e;

c) chegar a um consenso sobre quais são os elementos descritivos mínimos necessários para facilitar a recuperação da informação.

2.3 A Linguagem de Marcação Extendida (XML) e a Estrutura para Descrição de Recursos (RDF)

Em Unidades de Informação, atualmente temos o formato de intercâmbio MARC – *Machine Readable Cataloging*, para a descrição bibliográfica, respondendo às necessidades de informatização dos catálogos.

No entanto, esse formato não consegue suprir as necessidades de padronização e recuperação no contexto da Internet, pois não possui uma linguagem de fácil aplicação por qualquer utilizador, e, que, ao mesmo tempo possa ser interpretada pelos *browsers*.

A linguagem XML – *Extensible Markup Language*, pode ser aplicada para atender essa necessidade no contexto *WEB*, com o objetivo de facilitar a difusão da informação documental. Por possuir uma semântica própria, descreve a estrutura e o conteúdo do documento, não a sua formatação, sendo este o maior diferencial em relação à linguagem HTML, que apenas possibilita a formatação dos dados no que diz respeito a sua apresentação gráfica e não fornecendo nenhum conteúdo semântico.

As marcas descritoras do HTML são fixas, logo, as formatações dos documentos são limitadas. A passagem de documentos em formatos próprios para HTML, necessita de uma prévia conversão, em que habitualmente se perdem algumas características do texto. O utilizador está limitado a um conjunto de tipos de formatações específicas, sendo impossível, por exemplo, fazer aparecer fórmulas com símbolos matemáticos ou químicos em HTML – sem recurso a imagens anexas.

Ao utilizar a linguagem HTML, não é possível manter uma base de índices de documentos para depois fazer pesquisas temáticas. A linguagem XML pode suprimir estas limitações, pois ela possui grandes potencialidades para se tornar efetivamente, um padrão no contexto abordado.

Quando incluímos a linguagem XML à descrição de recursos Web, surgem *n* variáveis com as quais precisamos lidar, desde a questão da semântica e da sintaxe, bem como a questão da interoperabilidade, devido à grande variedade de metadados existentes.

> De forma a poder ser utilizado em larga escala, a DCMI (*Dublin Core* Metadata Initiative) optou por definir de forma ampla a semântica do DC, deixando as questões ligadas à sintaxe abertas e indefinidas. Esta é a razão pela qual o DC e o RDF (Resource Description Framework) combinam tão bem: o RDF traz as regras sintáticas nas quais o DC pode ser embebido. (BAPTISTA, 2001, p. 77).

O RDF "é uma aplicação recomendada pelo W3C (World Wide Web Consortium) para codificar, fazer o intercâmbio e reutilizar os dados normalizados" (HAROLD apud BAPTISTA, 2001, p. 80).

O RDF é uma base de processamento de metadados que promove a interoperabilidade entre diversas aplicações que variam entre páginas Web ou seu conteúdo, até documentos armazenados em um computador pessoal. É uma linguagem e uma gramática para definir a arquitetura da metainformação na Web, que permite a descrição de recursos de maneira estruturada. Seu principal objetivo é representar metadados sobre recursos web, como título, autor, data de modificação de uma página, etc. Esses metadados permitem a evolução da busca semântica na Web contribuindo, assim, para o desenvolvimento de ferramentas para descrição de recursos web.

3 RECURSOS ANALISADOS

Roram utilizados os seguintes recursos: RDF; XML; tags META, descritos a seguir.

3.1 Análise do Recurso Descritivo RDF

O recurso descritivo RDF é uma arquitetura de metadados desenvolvida para a utilização com a tecnologia XML, e que permite integrar o padrão DC a essa linguagem.

> O surgimento de diversos padrões de metadados originou um grande problema: incompatibilidade entre os padrões. O padrão DC, no contexto de bibliotecas digitais, foi uma das primeiras tentativas de se gerar um padrão de metadados que fosse comum a todos os outros padrões. Apesar de ser um padrão aberto, ele não resolve o problema visto a natureza heterogênea de cada solução. É neste contexto que surgem as arquiteturas genéricas de metadados como solução para atingir interoperabilidade entre informações descritas em diferentes padrões de metadados. As arquiteturas oferecem a flexibilidade necessária em ambientes heterogêneos, permitindo que recursos possam ser descritos seguindo diversos padrões, aproveitando assim o que cada um tem de melhor em termos de semântica descritiva. (MARINO, 2001b, p. 17)

A arquitetura de metadados RDF visa a prover a interoperabilidade entre fontes de dados que foram desenvolvidas sob diferentes perspectivas de organização dos dados. Essa integração é baseada em uma camada semântica e exige a construção de três modelos. O modelo conceitual que representa uma simples ontologia do domínio de conhecimento da aplicação. O modelo lógico que permite expressar o esquema de fontes de dados estruturadas e semi-estruturadas. E por último, o modelo de mapeamento, que define a relação entre os elementos dos modelos lógico e conceitual. (MARINO, 2001a)

De acordo com Harold (apud BAPTISTA, 2001, p. 80), a estrutura RDF utiliza *statements* para fazer as declarações sobre recursos,

eles podem ser vistos como "[...] um triplo composto por três elementos: recurso (sujeito), propriedade (predicado) e valor (objeto)". O *stament* RDF faz essas declarações *por meio de* de uma propriedade, retornando como resultado da aplicação um valor.

Lassila e Swick (apud BAPTISTA, 2001) mostram a especificação de modelo e sintaxe RDF como uma recomendação do *World Wide Web Consortium* (W3C), para representar, codificar e transportar metadados em ambiente *web* otimizando a interação cliente-servidor, tornando-os mais independentes.

> A escolha da RDF como tecnologia para a especificação da proposta se deve a três grandes motivações. A primeira é que RDF, quando serializada em um formato XML, torna-se bastante apropriada para representar fontes de dados estruturadas e semi-estruturadas. A segunda é que RDF permite expressar dado e metadado usando o mesmo formalismo, possibilitando uma navegação uniforme entre eles. A terceira e talvez a mais importante motivação é a expressividade do formalismo RDF, fornecendo o suporte para o mapeamento entre os diferentes esquemas. (MARINO, 2001a, p. 3)

Esse recurso permite, portanto, integrar os elementos DC em um documento XML, com a devida sintaxe, sem que com isso o Padrão DC perca sua flexibilidade.

Para fins de ilustração, será apresentado a seguir um fragmento de arquivo contendo a descrição física dos elementos usando RDF sobre XML:

```
<?xml version="1.0"?>
    <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:dcq="http://purl.org/dc/qualifiers/1.0/">
    <rdf:Description rdf:about="http://www.dublincore.org/">
    <dc:title xml:lang="en">Dublin Core Metadata Initiative (DCMI) Home
Page</dc:title>
    <dc:description xml:lang="en">The Dublin Core Metadata Initiative
is an open forum engaged in the development of interoperable  online
metadata standards that support a broad range of purposes and business
models. DCMI's activities include  consensus-driven working groups,
global conferences and workshops, standards liaison, and educational
efforts to  promote widespread acceptance of metadata standards and
practices.</dc:description>
    <dc:date xml:lang="en">2004-06-14</dc:date>
    <dc:format xml:lang="en">text/html</dc:format>
    <dc:contributor xml:lang="en">Dublin Core Metadata
Initiative</dc:contributor>
    <dc:language xml:lang="en">en</dc:language>

    <commercial:content-type xml:lang="en">text/html;
charset=iso-8859-1</commercial:content-type>
    </rdf:Description>
    </rdf:RDF>
```

Figura 1 – Fragmento de arquivo XML com elementos RDF

3.2 Análise do uso da tecnologia XML aplicada ao padrão DC

Metadado tem sido considerado um elemento fundamental no suporte a interoperabilidade de recursos que apresentam um alto grau de distribuição e heterogeneidade, em especial no contexto das aplicações científicas, uma vez que permite a conceitualização dos objetos normalmente complexos encontrados neste tipo de ambiente (MARINO, 2001b, p. 6)

A linguagem XML veio preencher uma lacuna nas necessidades de padronização e recuperação no contexto da Internet, com o objetivo de facilitar a difusão da informação documental. Por possuir uma semântica própria, descreve a estrutura e o conteúdo do documento, não a sua formatação, tornando-se revolucionário em relação à linguagem HTML.

A linguagem de marcação XML possibilita identificar diferenças semânticas entre ontologias específicas, ou seja, contextualiza a informação ou dados descritos em um documento *web*. (BAX, 2001).

O conjunto de descritores do *Dublin Core*, pode estar intrínsecos no próprio documento descrito por meio da linguagem HTML, XML e outras, ou, de acordo com o recurso necessário, a meta-informação pode estar separada do recurso utilizado para a catalogação. (DUVAL, 2002).

A flexibilidade e interoperabilidade inerentes à linguagem XML colocam-na em uma posição estratégica no contexto da recuperação da informação, já que é possível utilizá-la em conjunto com normas e protocolos específicos, como o Z39.50[1]. Além disso, permite que vários modelos de dados sejam intercambiados por instituições científicas.

O uso de *Document Type Definitions* (DTDs), é responsável por permitir essa interoperabilidade, já que é *por meio de* das DTDs que cada comunidade expressa seu contexto. Os *namespaces* da linguagem XML são outro fator importante nas aplicações para metadados, que podem identificar qual é o modelo utilizado para a descrição dos recursos, permitindo assim a integração do DC aos outros padrões.

A seguir apresenta-se um fragmento de texto de um arquivo em formato XML, contendo a descrição no padrão DC:

---

[1] O Z39.50 é um padrão internacional (ISO 23950) que define um protocolo de comunicação de um computador para outro que permite recuperar informações entre sistemas heterogêneos. Esse padrão torna possível que um usuário efetue uma busca e recupere informações de outro computador (que tenha o Z39.50 implementado) sem conhecer a sintaxe específica de outros sistemas. (LC, 2004)

```
<?xml version="1.0" ?>
  <metadata
    xmlns="http://www.ukoln.ac.uk/metadata/dcdot/"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://www.ukoln.ac.uk/metadata/dcdot/ http://www.ukoln.ac.uk/metadata/dcdot/dcdot.xsd"
    xmlns:dc="http://purl.org/dc/elements/1.1/">
  <dc:title xml:lang="pt">
    Portal On Line do Centro de Ciencias da Educacao
  </dc:title>
  <dc:creator>
    Centro de Ciencias da Educacao - CCE - Universidade do Estado de Santa Catarina - UDESC
  </dc:creator>
  <dc:subject xml:lang="pt">
    Centro de Ciencias da Educacao - CCE - Universidade do Estado de Santa Catarina - UDESC
    cursos universitarios florianopolis faculdade nivel superior mestrado doutorado pos-graduacao
    historia geografia pedagogia biblioteconomia
  </dc:subject>
  <dc:description xml:lang="pt">
    Portal On Line do Centro de Ciencias da Educacao - CCE - Universidade do Estado de Santa
    Catarina - UDESC
  </dc:description>
  <dc:publisher>
    Centro de Ciencias da Educacao - CCE - Universidade do Estado de Santa Catarina - UDESC
  </dc:publisher>
  <dc:contributor>
    HACK, Luciano
  </dc:contributor>
  <dc:contributor>
    AMORIM, Dorian
  </dc:contributor>
  <dc:contributor>
    ROSA, Silvana Bernardes
  </dc:contributor>
  <dc:date>
    2003
  </dc:date>
  <dc:type>
    Text
  </dc:type>
  <dc:format>
    text/html
  </dc:format>
  <dc:format>
    1002 bytes
  </dc:format>
  <dc:identifier>
    http://www.faed.udesc.br
  </dc:identifier>
  <dc:source>
    http://www.faed.udesc.br
  </dc:source>
  <dc:language>
    pt
  </dc:language>
  <dc:relation>
    http://www.udesc.br
  </dc:relation>
  <dc:coverage>
    SC-BR
  </dc:coverage>
  <dc:rights>
    Copyright - All rights reserved
  </dc:rights>
  </metadata>
```

Figura 2 - Implementação do registro DC em XML.

3.3 Implementação do DC em tags META

Documentos implementados em HTML podem conter na seção *header* as tags META, que normalmente são utilizadas para conter elementos descritivos pertinentes ao conteúdo da página ou portal.

A implementação do padrão DC em tags META pode ser realizada editando-as e inserindo as especificações e valores do padrão DC. O fragmento de uma página, apresentado a seguir, pode ilustrar melhor a sintaxe desta forma de implementação:

```
<link rel="schema.DC" href="http://purl.org/DC/elements/1.1/">
<META NAME="DC.title" CONTENT="Dublin Core Metadata Initiative (DCMI)
Home Page">
<META NAME="DC.description" CONTENT="The Dublin Core Metadata
Initiative is an open forum engaged in the development of interoperable
 online metadata standards that support a broad range of purposes and
business models. DCMI's activities include  consensus-driven working
groups, global conferences and workshops, standards liaison, and
educational efforts to  promote widespread acceptance of metadata
standards and practices.">
<META NAME="DC.date" CONTENT="2004-06-14">
<META NAME="DC.format" CONTENT="text/html">
<META NAME="DC.contributor" CONTENT="Dublin Core Metadata Initiative">
<META NAME="DC.language" CONTENT="en">
<META NAME="Content-Type" CONTENT="text/html; charset=iso-8859-1">
```

Figura 3 – Fragmento de implementação das tags META

3.4 Integrando as Tecnologias e Ferramentas

Durante a pesquisa foi objetivado o estudo das tecnologias apresentadas, sendo necessário, no entanto verificar como essas ferramentas se integram para atender a presente proposta.

Os 15 elementos DC são representados em um documento XML, por meio de seus *namespaces*, ao descrever-se o recurso. As DTDs necessárias para definir os elementos DC encontram-se disponíveis na página Internet do DC *Metadata Initiative* (DCMI), bem como exemplos de estruturas RDF para a validação XML.

Estes elementos do padrão podem ser representados nas tags META de um documento HTML, que também podem ser construídos com apoio de softwares de livre distribuição ou comerciais, disponíveis pelo endereço http://www.dublincore.org .

Uma vez disponibilizados por meio de um servidor web, os recursos descritos com o padrão DC estão prontos para que o software indexador e de busca (*webspider*) possa coletar e armazenar os metadados produzidos em um banco de dados SQL, ou um sistema de índices na forma

de arquivos invertidos, dependendo do aplicativo utilizado para este fim, para serem acessados pelo sistema de recuperação.

O software indexador localiza as *tags* XML que representam os *namespaces* XML dos elementos DC, ou as tags META, no caso de arquivos HTML, identificando em qual campo do Banco de Dados deve ser depositado cada metadado ou que metadado deverá ser incorporado ao índice. Em um segundo momento, os índices são gerados, para serem utilizados posteriormente pelo motor de busca do software.

A combinação de softwares livres de uso amplamente difundido, com o padrão DC é uma boa opção para um projeto de recuperação de documentos da internet, em virtude de tornar viável a implantação de bases de dados de recursos de internet de forma tratada, com recursos acessíveis e com baixo custo de implantação e manutenção.

4 RESULTADOS

Essa análise conclui que no contexto e no momento atual da pesquisa, o *HtDig* continua sendo o aplicativo mais adequado para implementação de testes, pela sua simplicidade de implementação. Entretanto, verificou-se também que é possível utilizar outros mecanismos de indexação e busca, com o mesmo objetivo. Outras linguagens de programação para Internet também podem permitir o desenvolvimento de aplicações semelhantes, como por exemplo a linguagem JAVA e PHP, mas não foram exploradas no contexto desta pesquisa.

4.1 O Modelo proposto

O modelo proposto é essencialmente a combinação de recursos e tecnologias para implantação de um sistema de recuperação de

páginas da internet. Podemos situá-lo em um nível de abstração suficiente para ser independente de tecnologias, e ao mesmo tempo, constituir-se como um norteador para elaboração de um sistema de recuperação de documentos descritos pela aplicação de metadados DC.

O diagrama a seguir representa o modelo, composto se segue:

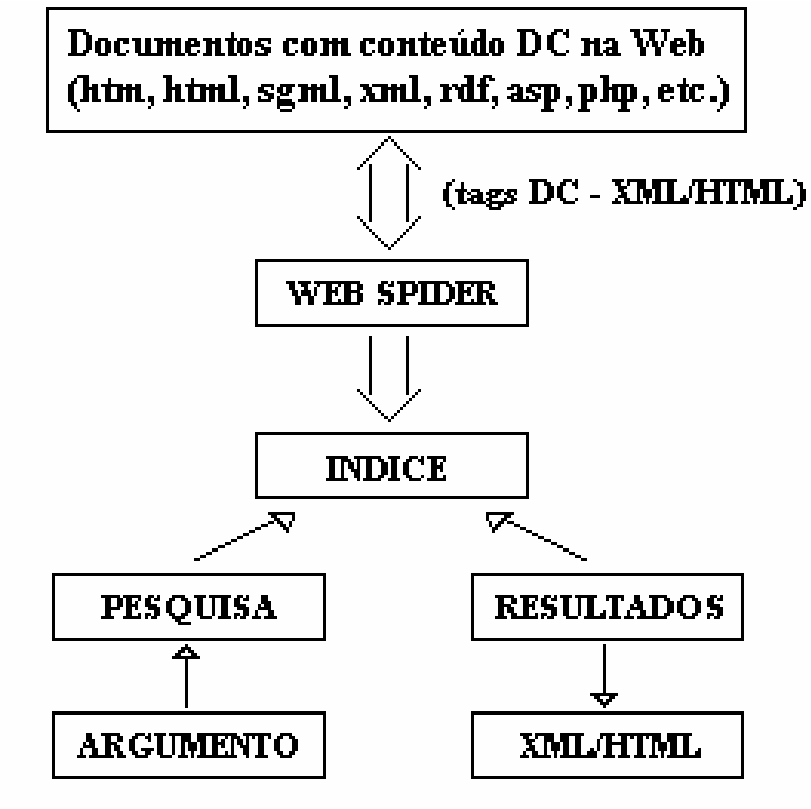

Figura 4 – Diagrama estrutural e funcional do modelo proposto

Este diagrama pode ser assim descrito:

a) preparação técnica do documento, consiste em descrever o documento objeto em um documento XML (ou HTML, se for o caso). Deve ser efetuado pela instituição que disponibiliza o documento ou descrição na *web*.

b) localização dos metadados, *por meio de* um *software* de indexação e busca (webspider), adequadamente configurado, detecta-se a informação solicitada.

c) indexação, realizado pelo mesmo software, processo no qual as palavras-chave das *tags* especificadas na configuração são coletadas, formando um índice alfabético que facilite a recuperação dos metadados.

d) pesquisa ou busca, por meio de uma interface disponível pelo software de indexação e busca, no qual termos de busca fornecidos pelo usuário são comparados pelo motor de busca com os índices, produzidos pelo processamento dos registros em DC.

Utilizou-se como plataforma de testes um microcomputador com sistema operacional *Linux*, servidor web Apache e motor de busca *HtDig*.

Para efetuar essa operação, é necessário que o webspider (*HtDig*) saiba o que procurar, nesse caso, os elementos DC em cada arquivo HTML, por meio da configuração de um arquivo de diretrizes, utilizado no momento da criação dos índices.

Ao receber uma solicitação de busca, o motor de busca deve ser capaz de identificar qual elemento DC o usuário está solicitando e comparar com um índice correspondente. O motor de busca deve ainda apresentar os resultados de uma consulta. Esse resultado é apresentado sob a forma de uma página criada dinamicamente, em HTML, com os resultados da busca.

4.2 Estrutura dos aplicativos envolvidos

A implementação de um sistema de recuperação de itens documentários da internet descritos com o padrão DC pode ser esquematizada como se segue:

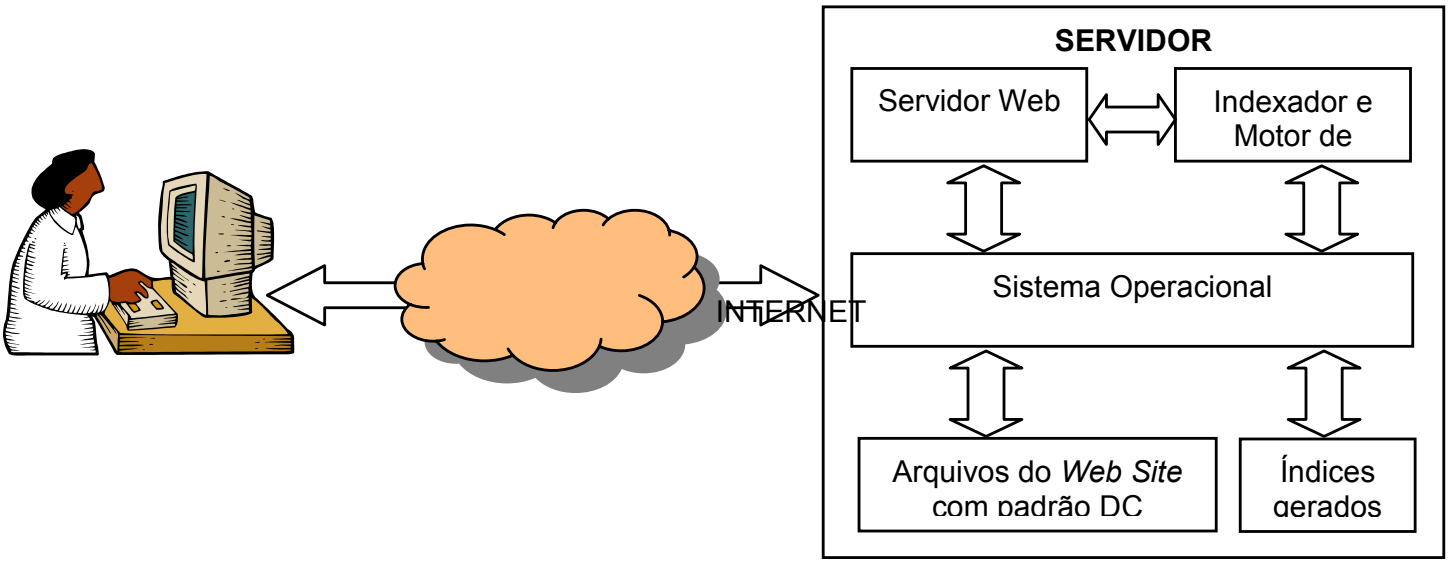

Figura 5 - Estrutura das aplicações em um servidor para Recuperação de Informação com DC

As políticas de seleção das plataformas, sistemas operacionais e tecnologias disponíveis são definidas, basicamente, em função de custos e capacidade de manutenção dos sistemas. Para essa pesquisa foi a política de *software* livre, utilizando um PC com *Linux* Red Hat, servidor Apache, *webspider HtDig*, e páginas HTML com padrão DC implementado por meio das tags META. Esta opção foi feita pela simplicidade de manutenção em montagem da plataforma de testes.

Em outros casos pode-se utilizar tecnologias compatíveis, substituindo a plataforma, o servidor e o *webspider* já que o foco da aplicação é o Padrão DC. É possível até desenvolver um mecanismo de busca particular ou proprietário, implementado com uma linguagem de programação, como PHP ou JAVA, já que ambas possuem processadores de arquivos que seguem o padrão XML, o que permite implementar o mecanismo de busca.

A seguir, um quadro comparativo, com alguns motores de busca com os quais pode-se desenvolver sistemas de recuperação de informação com aplicação do padrão DC:

Característica	Htdig	Swish-e	Lucene	Aspseek
Compatível com ambiente UNIX ou Linux	SIM	SIM	SIM	SIM
Compatível com ambiente Windows	SIM	NÃO	SIM	NAO
Suporte a meta tags definidas pelo usuário (Dublin Core)	SIM	Somente se definidas nas tags META dos arquivos	Por módulo extrator adequado	NÃO
Operadores booleanos	SIM	SIM	SIM	SIM
Tratamento de arquivos em outros formatos (pdf, doc, ppt)	Por meio de filtros	Por meio de filtros	Por meio de filtros	Por meio de conversores externos
Site da ferramenta	http://www.htdig.org	http://swish-e.org	http://jakarta.apache.org/lucene	http://www.aspseek.org

Quadro 1 - Comparação entre mecanismos de busca (CRUZ, 2003, p.2)

O equipamento recomendado pode ser um computador com um processador Pentium 4 ou superior ou equivalente, com 512Mb de RAM, HD de 40Gb, e conexão rápida com a internet (cabo, radio, adsl, etc.). Esta configuração poderá ser melhorada, conforme as necessidades de acesso aos recursos.

5 CONSIDERAÇÕES FINAIS

A presente pesquisa apresenta uma revisão de literatura sobre os diversos conceitos de metadados, do padrão Dublin Core e sua utilização, bem como das ferramentas disponíveis para Recuperação da Informação na Internet e suas aplicações quanto à descrição de dados on-line, permitindo elaborar um modelo estruturado para implementação de um software capaz de efetuar pesquisas na Web por meio de metadados.

Ao estudar os conceitos que envolvem os metadados, foi possível compreender sua utilidade e importância para a Ciência da Informação e para a Recuperação da Informação na Internet, e a necessidade de integrar interdependentemente cada uma das ferramentas ao padrão Dublin Core com o objetivo de subsidiar um modelo de um *site* ma de busca para páginas da Internet.

A partir de análises baseadas na literatura e por meio de testes realizados com as ferramentas tecnológicas disponíveis durante o desenvolvimento da pesquisa propôs-se um modelo funcional com o objetivo de subsidiar a implementação de recursos de um software de forma a dispor um mecanismo de busca capaz de localizar e manter uma base de referências de documentos que aplicam o padrão DC como recurso descritivo.

O modelo proposto é essencialmente a combinação de recursos e tecnologias para implantação de um sistema de recuperação de páginas da internet. Podemos situa-lo em um nível de abstração suficiente para ser independente de tecnologias, e ao mesmo tempo, constituir-se como um norteador para elaboração de um sistema de recuperação de documentos descritos pela aplicação de metadados DC.

Utilizou-se como plataforma de testes um microcomputador com sistema operacional *Linux*, servidor web Apache e motor de busca *HtDig*.

O papel do motor de busca *HtDig'* é criar os índices e realizar as buscas, servindo como uma ponte entre os dados semi-estruturados por meio de do padrão DC nas páginas organizadas no servidor web. Estas por sua vez contêm em suas *tags* META os elementos descritivos segundo o padrão DC.

Verificou-se que o padrão Dublin Core é ainda pouco utilizado no Brasil e, portanto, faz-se necessário divulgar e incentivar seu uso como estratégia para garantir a continuidade do mesmo. Entre os poucos casos encontrados na literatura abordando a utilização ou contribuição do padrão DC no Brasil, pode-se citar:

a. o caso da Embrapa Informática Agropecuária (SOUZA, 2000), que utilizou o padrão DC modificado (adaptando-o à sua necessidade) para o Banco de Imagem Rural Mídia (BI-RM);
b. o Projeto Biblioteca Digital Temática do Empreendedor do SEBRAE Paraíba (GASPAR, 2004), que utiliza o padrão DC para alimentar a Base de Dados da Biblioteca Digital, motivado por sua simplicidade, que permite a usuários leigos descrever os recursos necessários;
c. e o projeto "Casa de Ferreiro" da Universidade Católica de Pelotas (BONATTO, 2003), que utiliza o padrão DC como descritor para construção do Banco de Dados utilizado pelo motor de busca desenvolvido pelo projeto, apresentando novamente as aplicações do Dublin Core para o usuário final no Brasil.

Essas experiências demonstram claramente as possibilidades de implementação do Dublin Core no contexto da sociedade brasileira.

Com as reflexões baseadas nas análises realizadas, propomos algumas aplicações práticas, empregando as tecnologias disponíveis, tais como: a utilização de técnicas de Representação Descritiva; a conversão de

registros MARC21 para DC; e, a verificação do uso de padrões de metadados geoespaciais.

Verificou-se que é possível utilizar esses conhecimentos para a elaboração de projetos de Recuperação da Informação e interfaces de pesquisa para efetuar pesquisas baseadas no padrão DC, ou outros padrões de metadados.

Considerou-se que é pertinente implementar um software com base no modelo apresentado com o intuito de motivar a utilização do padrão Dublin Core, bem como dos metadados e das tecnologias atuais, e aplicar programas de divulgação dos mesmos para efetivamente facilitar a recuperação e acesso de informações em rede, seja interna (Intranet) ou externa (Internet, Extranet, WAN, etc.).

REFERÊNCIAS

ALVARENGA, Lídia. Representação do conhecimento na perspectiva da Ciência da informação em tempo e espaços digitais. *Encontros Bibli,* Florianópolis, n. 15, 2003. Disponível em: <http://www.encontros-bibli.ufsc.br/Edição_15/alvarenga_representação.pdf>. Acesso em: 12 jan. 2004.

ALVES, Rachel Cristina Vesú. *Análise dos Padrões de Descrição das Informações para Organização de Documentos Eletrônicos*. 2002. 104 f. Trabalho de Conclusão de Curso (Bacharelado) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília.

BAPTISTA, Ana Alice; MACHADO, Altamiro Barbosa. Um gato preto num quarto escuro: falando sobre metadados. *Revista de Biblioteconomia de Brasília,* Brasília, v. 25, n. 1, p. 77-90, maio/jun 2001.

BAX, Marcelo Peixoto. Introdução às Linguagens de Marcas. *Ciência da Informação*, Brasília, v. 30, n. 1, p. 32-38, jan./abr. 2001.

BONATTO, Daniel Torres. *Casa de Ferreiro.* ago. 2003. 74 f. Relatório (pesquisa). - Escola de Informática – Universidade Católica de Pelotas,

Pelotas, 2003. – Disponível em: <http://gpia.ucpel.tche.br/cva/Casa_de_Ferreiro/Casa_de_Ferreiro.doc>. Acesso em: 22 jul. 2004.

CRUZ, Sérgio A. B. da. Implantação de Um Serviço de Busca em Site da WWW. *Comunicado Técnico*, Campinas(SP), n. 50, novembro 2003. Disponível em:<http://www.cnptia.embrapa.br/modules/tinycontent3/content/2003/comtec50.pdf>. Acesso em: 20 jul. 2004.

DESAI, B. C. Supporting Discovery in virtual libraries. *Journal of the American Society for Information Science,* v. 48, n. 3, p. 190-204, 1997.

DUVAL, Erik et al. Metadata Principles and Practicalities. *D-Lib Magazine,* v. 8, n. 4, p. 1-15, abr., 2002. ISSN 1082-9873. Disponível em: <http://www.dlib.org/dlib/april02/weibel/04weibel.html>. Acesso em: 15 nov. 2002.

FROST, G. *User's guide to Dublin Core* document description. 26 jul. 1997. Disponível em: http://www.valdosta.edu/~grost/. Acesso em: 2002.

GARCIA, S. S. *Metadados para documentação e recuperação de imagens***.** 1999. 138 f. Dissertação (Mestrado). - Instituto Militar de Engenharia – IME, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 1999.

GASPAR, C. de O.; MOTA, A. R. S.; PAULO, V. V. *Biblioteca Temática do Empreendedor:* relato da experiência do SEBRAE Paraíba na construção de uma biblioteca digital. Trabalho apresentado no II Simpósio Internacional de Bibliotecas Digitais, Campinas, 2004. Disponível em: <http://libdigi.unicamp.br/document/?view=8281>. Acesso em: 22 jul. 2004.

GOMES, G. R. R.; MELO, R. N.; CÔRTES, S. C. *Uma arquitetura de informática para integração de sistemas de bibliotecas na Internet.* PUC-Rio.[2001]. Trabalho baseado na dissertação de mestrado "Um ambiente para integração de dados bibliográficos baseados em mediadores".

GRÁCIO, José Carlos Abbud. *Metadados para a Descrição de Recursos da Internet:* o padrão *Dublin Core*, aplicações e a questão da interoperabilidade. 2002. 104 f. Dissertação (Mestrado). - Programa de Pós-Graduação em Ciência da Informação. Universidade Estadual Paulista, Marília, 2002

LIBRARY OF CONGRESS (LC). *WWW/Z39.50 Gateway*. Disponível em: <http://lcweb.loc.gov/z3950/gateway.html#about>. Acesso em: 30 jun. 2004.

MARINO, Maria Teresa. Capítulo I: introdução. In: ___. *Integração de Informações em Ambientes Científicos na Web***:** uma abordagem baseada na arquitetura RDF.2001. Dissertação (Mestrado). - Programa de Pós-Graduação em Informática. Universidade Federal do Rio de Janeiro. Rio de Janeiro: UFRJ, 2001a.

MARINO, Maria Teresa. Capítulo I: introdução. In: ___. *Integração de Informações em Ambientes Científicos na Web***:** uma abordagem baseada na arquitetura RDF.2001. Dissertação (Mestrado). - Programa de Pós-Graduação em Informática. Universidade Federal do Rio de Janeiro. Rio de Janeiro: UFRJ, 2001b.

PEREIRA, A V. G.; TAVARES, G. C. O; MARTINS, J. M. P. N.; COELHO, M. P. S. *Metadados***:** sistemas de informação geográfica. Disponível em: <http://www.isa.utl.pt/dm/sig/sig20002001/TemaMetadados/trabalho.htm>. Acesso em: 20 nov. 2002.

SOUZA, M. I. F.; VENDRUSCULO, L. G.; MELO, G. C. Metadados para a descrição de recursos de informação eletrônica: utilização do padrão *Dublin Core*. *Ciência da Informação,* v. 29, n. 1, p. 93-102, jan./abr. 2000.

SOUZA, T. B. de; CATARINO, M. E.; SANTOS, P. C. dos. Metadados: catalogando dados na Internet. *Transinformação,* v. 9, n. 2, maio/ago. 1997. Disponível em: <http://www.biblioestudantes.hpg.ig.com.br/146.htm>. Acesso em: 05 mar. 2004.

_____

METADATA FOR INTERNET RESOURCES DESCRIPTION: new technologies developed for the Dublin Core standard and its use

Abstract: *This research has as objectives the development knowledge in the production of software what the information of recuperation through of the available technology for Internet (XML, JAVA, or PHP, RDF, Dublin Core), the analysis the working of the standard of metadata Dublin Core and the structure of descriptors RDF (Resource Description Framework), that begin of the base of sustentation in the development from that software. The identify the languages that's begin used in the creation of the software (XML, JAVA, or PHP), to subsidize of development of same, to analyse the syntax and the use of the available technology, in application of the Dublin core standard for information of recuperation for middle of the meta-tags of*

the HTML, XML technology, and of the structure of the description of RDF resources. The research was literature review and considered the availability from these tools for common user in the network. As conclusion verified that analyses and studies of the metadata's and of the available technology permitted to describe as to construct one interface of search in the standard Dublin core and the XML technology, contributing thus, for practices application and advantages that the study of Metadata proportionate as such for academic middle, and also for the society.

***Keywords**: Information Retrieval; Metadata; Dublin Core; Search Engines*

_____

MSc. Ana Maria Pereira
Docente de Biblioteconomia do Centro de Ciências da Educação – UDESC
E-mail: pereiraana_maria@hotmail.com

MSc. Divino Ignácio Ribeiro Júnior
Docente de Biblioteconomia do Centro de Ciências da Educação – UDESC
E-mail: f2dir@udesc.br

Guilherme Luiz Cintra Neves
Acadêmico de Biblioteconomia do Centro de Ciências da Educação – UDESC
Bolsista de Iniciação Cientifica – PROBIC/UDESC
E-mail : guilhermecneves@ibestvip.com.br