Metadados no Contexto da Web Semântica

Alexandre Mori *C. L. de Carvalho*

Technical Report - RT-INF_002-04 - Relatório Técnico
November - 2004 - Novembro

Instituto de Informática
Universidade Federal de Goiás
www.inf.ufg.br

Metadados no Contexto da Web Semântica

Alexandre Mori [*] xmori@bol.com.br

Cedric Luiz de Carvalho [†] cedric@inf.ufg.br

Abstract. *This work discuss the importance of metadata, specially in the Semantic Web Context. It is considered some patterns, applications, the necessity of using metadata and its relation with other technologies, such as XML, for instance.*

Keywords: Semantic Web, Metadata, XML, Metadata Patterns.

Resumo. *Neste trabalho é discutida a importância dos metadados, especialmente no contexto da Web Semântica , nos dias atuais. São considerados alguns padrões, aplicações, qual a necessidade do uso de metadados e sua relação com outras tecnologias como XML, por exemplo.*

Palavras-Chave: *Web* Semântica, metadados, XML, padrões de metadados.

1 Introdução

Metadados são comumente definidos como dados sobre dados. Quando se trata deste assunto, esta é a definição mais simples e comumente utilizada. A palavra originou-se do latim *metá* que significa "além", "através de" ou "sobre´´, isto é, dados sobre outros dados.

Muitos autores têm sua própria definição, outros dizem que, por não possuir um único objetivo, metadados não têm uma definição ampla o bastante para abranger todo seu significado. Para Codd [3], por exemplo, "metadados consistem de dados que descrevem todos os outros dados em um banco de dados ...". Ikematu [14] selecionou algumas outras definições sobre metadados:

- Metadados são dados que descrevem atributos de um recurso. Eles suportam um número de funções: localização, descoberta, documentação, avaliação, seleção, etc.
- Metadados fornecem o contexto para entender os dados através do tempo.
- Metadados são dados associados com objetos que ajudam seus usuários potenciais a terem vantagem completa do conhecimento da sua existência ou características.
- Metadados são o instrumental para transformar dados brutos em conhecimento.

[*] Aluno do Curso de Especialização em Sistemas de Informação / GEApIS/INF/UFG
[†] Orientador - GEApIS/INF/UFG

Poderiam ser colocadas aqui inúmeras outras definições encontradas em vários artigos e trabalhos. Mas, como se vê, não há um consenso quanto à definição de metadados.

A seguir se discute mais detalhadamente temas como padrões de metadados e sua organização. Para iluminar a discussão, são apresentados também alguns exemplos de sua aplicação.

O restante deste texto está estruturado da seguinte forma: na seção 2, é discutido o que vem a ser a *Web Semântica*; na seção 3, como os metadados se incluem no contexto da *Web* ; na seção 4, se comenta um pouco sobre XML; na seção 5, são mostrados alguns dos padrões de metadados, bem como são detalhados os padrões MARC e Dublin Core; na seção 6, são discutidos tipos de metadados e, por fim, algumas considerações finais são tecidas na seção 7.

2 A *Web Semântica*

A Internet, mais epecificamente a sua porção multimídia, a *World Wide Web*, ou simplesmente *Web*, armazena uma enorme quantidade de dados espalhados pelo mundo todo. Tamanha é sua heterogeneidade que não é difícil de se entender o porquê da dificuldade de se encontrar informações acerca de algo específico.

Os dados disponibilizados na Internet foram modelados de maneira a apresentarem informações às pessoas. Ela cresceu como um meio de comunicação entre humanos e não direcionada às máquinas e computadores. Esta formatação dos dados, direcionada para a sua apresentação, dificulta a sua manipulação automatizada por meio de máquinas (computadores), uma vez que estas nem sempre são capazes de identificar a semântica associada a estes dados.

Neste ambiente, as buscas, normalmente, são realizadas a partir de palavras-chaves fornecidas a mecanismos de buscas, como o Google [1], por exemplo, que percorrem a rede tentando encontrar ocorrências delas. Por exemplo, se for pedido que o computador procure por “vinho”, ele, provavelmente, rastreará textos que contenham esta palavra. A pobreza do resultado produzido é conseqüência da incapacidade de se identificar o real significado dos dados acessíveis. O sítio da marca “X” de vinhos poderia não ser recuperado, sendo que, talvez, fosse ele o objetivo do usuário. Não seria, também, surpreendente se outros tantos sítios relevantes ficassem fora do resultado da pesquisa.

Utilizando-se de metadados, os sítios poderiam ser rotulados da forma pela qual desejassem ser encontrados. Como exemplo, pode-se imaginar uma adega com todas suas garrafas de vinho sem rótulos, apenas as quinhentas garrafas e os seus respectivos conteúdos. O dono da adega poderia saber o conteúdo de cada garrafa sem precisar de nenhum outro dado, pois ele organizara sua adega de sua maneira particular.

Mas, se uma outra pessoa que ali chegasse procurasse um vinho tinto de mesa produzido em Portugal, talvez demorasse bastante para encontrar tal produto. Ou, na pior das hipóteses, não o encontraria.

Os rótulos fariam o papel de metadados no contexto da *Web Semântica*. O vinho em si seriam os dados (o conteúdo) de algum sítio (garrafa). Sem os metadados (rótulo das garrafas) a procura seria menos eficiente. A adega faria, em menor proporção, o papel da Internet.

Até agora se discutiu a importância dos metadados, mas não foi comentado nada sobre a *Web Semântica*. Ora, esta baseia-se, essencialmente, nos metadados. Ela foi idealizada compartilhando a idéia de dados sobre dados.

Com os dados “rotulados”, computadores podem obter informações e disponibilizá-las ao seu usuário. Podem encontrar dados mais rápida e precisamente. Este é o objetivo da *Web Semântica*: prover um meio comum que permita que dados sejam compartilhados e reutilizados entre aplicações, empresas e comunidades [24].

A *Web Semântica* é considerada uma extensão da *Web* atual onde o significado da informação é bem definido, melhorando o trabalho cooperativo entre computadores e pessoas [24]. Assim é definida a *Web Semântica*: a *Web* atual estruturada de forma a permitir que máquinas consigam captar o significado do conteúdo de cada recurso. Deve-se entender "recurso" como qualquer fonte de informação.

Mas, como se pode adicionar significado aos dados? De que maneira a Internet pode ser reestruturada? Na próxima seção estas questões são discutidas.

3 Metadados na *Web*

A linguagem HTML (*Hypertext Markup Language*), ou lingugem de marcação de hipertextos, é bastante conhecida. Ela foi criada por Tim Berners-Lee[1] no início da década de 90 com o objetivo de facilitar a divulgação de informações pela *Web*. A HTML é um subconjunto da SGML[2] (*Standard Generalized Markup Language*) ou linguagem de marcação generalizada padrão.

O intuito era criar uma linguagem que pudesse ser "entendida" por todos os computadores. Uma que, independente de plataforma, fosse capaz de exibir informações criadas a partir de outros computadores.

Desde então, a HTML passou por várias revisões até chegar em sua última especificação: a 4.01 (recomendada pelo W3C em 29 de dezembro de 1.999).

Um arquivo no formato HTML é construído por textos e *tags* (veja o Exemplo 1). As *tags* são marcações dentro do arquivo que formatam o texto. Elas são formadas por um sinal "<", o identificador da formatação e o sinal ">". Veja a *tag* <**b**>(linha 14) : ela indica que o texto será evidenciado (negrito). Note que, ao final do texto a ser apresentado, há uma outra *tag* finalizando a primeira (****), diferenciando-se desta somente por conter uma barra ("/") após o sinal "<". Algumas *tags* são, até certo ponto, intuitivas (*tag* <**head**>(linha 2) e *tag* <**body**>(linha 11)), outras, porém, nem tanto (*tag* <**b**>(linha 14)).

[1] Tim Berners-Lee também é fundador do W3C e idealizador da *Web Semântica*

[2] A SGML foi criada entre 1960 e 1970. Sua especificação é bastante extensa, o que a torna consideravelmente complexa.

Exemplo 1 – Exemplo de arquivo HTML.

```
<html>
<head>
  <title>Cedric Luiz de Carvalho</title>
  <meta http-equiv="Content-Language" content="pt-br">
  <meta http-equiv="Content-Type" content="text/html;
                                   charset=windows-1252">
  <meta content="Microsoft FrontPage 4.0" name="GENERATOR">
  <meta content="FrontPage.Editor.Document" name="ProgId">
  <meta name="Microsoft Theme" content="none">
</head>
<body>
    <p>
    <a href="file://ZEUS/cedric/public_html/index.htm">
    <b>Portuguese</b></a>

    .
    .
    .
```

O que se pode notar é que há uma quantidade considerável de *tags*. Na HTML, estas *tags* são para formatar o texto, isto é, manipular a forma de apresentação. Tudo gira em torno da exibição dos dados.

As *tags* da HTML não deixam de ser metadados, pois são dados sobre dados, isto é, ao se inserir um texto "teste" entre as tags <**b**> e </**b**>irá ocorrer a formatação da palavra dada. Ela ficará negritada pois estas *tags*, em HTML, fazem esta alteração no texto que está entre elas.

Observa-se que as *tags* – não informam nada sobre o dado em si. A palavra "teste" continua sem um significado específico. Sabe-se o que esta palavra significa, mas, a que tipo de teste ela se refere?

Para se atribuir um significado específico a ela, pode-se construir a seguinte sentença:

<exemplo>teste</exemplo>

Assim, a palavra teste estaria servindo como um exemplo. Mas, exemplo de quê? Pode-se acrescentar outras sentenças:

<marcação>
<exemplo>teste</exemplo>
</marcação>

Observa-se que a sentença original agora está entre as *tags* **<marcação>** e **</marcação>**. Isto significa que a palavra "teste" é um exemplo de marcação. Assim pode-se acresentar quantas *tags* forem necessárias para atribuir um determinado nível de detalhamento dos dados.

Mas, o programador HTML irá observar que estas *tags* não existem em HTML. De fato, as marcações aqui utilizadas foram somente para atribuir semântica ao dado "teste". O objetivo não era formatar a palavra.

Novamente, voltando-se ao exemplo da adega, pode-se imaginar que exista somente um conjunto limitado de rótulos para identificar todos os vinhos. A utilização da XML (*Extensible Markup Language*) permite muito mais flexibilidade. Com ela, pode-se criar *tags* personalizadas. Na seção a seguir, se discute um pouco mais sobre XML.

4 Linguagem de marcação extensível (XML)

A XML [21] é, atualmente, o dialeto SGML mais simples criado para interoperabilidade entre várias plataformas e entidades. Deve-se observar que a XML também é um subconjunto da SGML, e por isto, tudo que é escrito em XML, teoricamente, também pode ser escrito em SGML.

Esta linguagem de marcação extensível é um padrão aberto que o W3C (*World Wide Web Consortium* [23]) projetou como um formato de dados para intercâmbio de documentos estruturados na *Web*. Ele estende suas opções de marcação, permitindo ao desenvolvedor definir seus próprios metadados quando a HTML não lhe atender.

"Dados estruturados" estão intimamente relacionados com coisas tais como folhas de cálculo, lista de endereços, parâmetros de configuração, transações financeiras, desenhos técnicos, etc.. Programas que produzem tais dados, freqüentemente, os armazenam em discos, utilizando um formato binário ou formato texto. Este último permite que, se necessário, os dados possam ser verificados, mesmo sem a utilização do programa que o produziu.

XML é uma coleção de regras, diretivas e convenções para projetar formatos textuais para dados. Os arquivos produzidos devem ser fáceis de serem gerados e também lidos por um computador. Esta geração não deve produzir dados ambíguos e deve evitar problemas comums, tais como falta de extensibilidade, carência de suporte à internacionalização e dependência de plataforma [25].

Considerando-se o Exemplo 2, observa-se que as *tags* que nele aparecem são marcações relativamente coerentes com o conteúdo dos dados associadas a elas. Por exemplo, **<titulo>** (linha 2) demonstra que ali inciar-se-á o título do documento, em **<resumo>** (linha 8) inciar-se-á o resumo do documento e em **<data>** (linha 22) a data de sua publicação.

Os metadados aí utilizados fornecem um significado aos dados primários. O que se segue após a **<titulo>** (linha 2) é "Recuperação de Informações Expressas por Metadados Através do Uso de Agentes Inteligentes", que é o dado propriamente dito (dado primário). Este dado não tem sentido sem as *tags* **<titulo>** e **</titulo>**, as quais marcam o texto, associando a ele uma semântica bem definida. Da mesma forma, as demais *tags*. Isto significa que estas *tags* representam dados sobre os dados primários, ou seja, metadados.

A XML possui muitos recursos, o que a torna completa para os propósitos já mencionados neste texto. Pode-se encontrar sua última especificação (04 de fevereiro de 2004) em [21]. Existem ainda os recursos dos DTDs (*Document Type Definition*) [21] , XSLT (*XML Stylesheet Language Transformations*) [22] , XMI (*XML Metadata Interchange*)[11] que complementam esta tecnologia.

Pode-se organizar todos os arquivos no formato XML, preenchê-los de metadados e publicá-los na Internet. Tem-se então a *Web Semântica* construída. Mas, de que forma se deve inserir os metadados em um arquivo que descreve um livro, por exemplo? Deve-se usar a *tag* **<title>** ou **<titulo>**? Quais os metadados necessários para suprir todas as informações que serão solicitadas? Estes questionamentos podem ser respondidos a partir da discussão a seguir.

Exemplo 2 – Exemplo de arquivo XML.

```
<documento>
        <titulo>
                Recuperação de Informações Expressas por
                Metadados Através do Uso de Agentes Inteligentes
        </titulo>
        <autores>Mateus Ricardo Provensi</autores>
        <orientador> Cedric Luiz de Carvalho </orientador >
        <resumo>
                Quando se efetua uma pesquisa na Web
                utilizando os tradicionais motores de
                busca, são obtidas muitas páginas que não
                são relevantes aos interesses dos
                usuários o que torna o acesso a
                informação uma tarefa difícil e
                demorada...
        </resumo>
        <palavras-chaves>
                Web Semática, Metadados, XML, RDF,
                Ontologias, Agentes e Acesso à
                Informação
        </palavras-chaves>
        <data>22/10/2003</data>
        <tipo>PFC</tipo>
        <formato>DOC</formato>
        <idioma>Português</idioma>
        <url>http://www.inf.ufg.br/~mateus/Relatorio2.doc</url>
 </documento>
```

5 Padrões de metadados

A finalidade principal dos metadados é documentar e organizar, de forma estruturada, os dados das organizações, com o objetivo de minimizar duplicação de esforços e facilitar a manutenção dos dados.

As corporações necessitam de um maior controle de seus dados, precisam conhecer melhor o conteúdo e a qualidade dos mesmos de forma rápida, automatizada e eficiente. Um outro motivo importante para se estabelecer padrões é a necessidade de disseminação da informação e o acesso à informação de propriedade de outras organizações.

Os padrões de metadados têm como função fornecer as definições e formar uma rede para automatizar registros de propriedades e dados cadastrais de uma forma padronizada e consistente.

Existem muitos padrões de metadados, que se diversificam em função de sua área de aplicação. Entre estes formatos pode-se destacar:

- *Government Information Locator Service* (GILS) − informações governamentais [9];
- *Federal Data Geographic Commitee* (FGDC) − descrição de dados geo-espaciais [4];
- *Machine Readable Cataloging Record* (MARC) − catalogação bibliográfica [18];

- *Dublin Core* (DC) − dados sobre páginas da *Web* [5];
- *Consortium for the Interchange of Museum Information* (CIMI) − Informações sobre Museus [2];
- *Directory Interchange Format* (DIF) − padrão para criar entradas de diretórios que descrevem um grupo de dados [8];
- *Meta Data Interchange Specification* (MDIS) − padrão para troca de metadados entre ferramentas da Tecnologia de Informação [16];
- *Open Information Model* (OIM) − conjunto de especificações para facilitar o compartilhamento e reuso no desenvolvimento de aplicações e *data warehouse* [6];
- *Common Warehouse Meta Model* (CWM) − padrão para troca de informações entre esquemas de banco de dados e *data warehouse*[10].

Nas duas subseções seguintes são apresentados mais detalhadamente dois padrões para metadados. O primeiro, o MARC [18] , é aplicado em catalogação bibliográfica, e o segundo, o Dublin Core [5] , é usado na descrição de recursos eletrônicos.

5.1 *Machine-Readable Cataloging Record* (MARC)

Nas grandes bibliotecas, obras literárias são catalogadas em fichas catalográficas. Uma ficha catalográfica consiste na catalogação do livro antes de sua publicação, identificando a obra de forma permanente e padronizada. Por permitir a identificação de um livro nele próprio, auxilia as bibliotecas no processo de catalogação de livros, facilita o registro bibliográfico, propicia a uniformização da citação bibliográfica, permite às editoras a organização de arquivos e, por fim, propicia informações concisas sobre a matéria abordada no livro, facilitando seu agrupamento por assunto e favorecendo sua veiculação. Um exemplo de ficha catalográfica pode ser visto na Figura 1.

A informação, tradicionalmente contida em uma ficha catalográfica, pode ser disponibilizada em formato eletrônico, em um registro bibliográfico (*Cataloging Record*). Este registro pode, obviamente, ser lido por máquinas - computadores (*Machine-Readable*).

O formato MARC é uma maneira de se registrar dados e metadados de maneira que máquinas possam lê-los. As fichas catalográficas tradicionais utilizam, normalmente, mas não necessariamente nesta ordem, os seguintes conjuntos de informações:

1. **Descrição**: Esta descrição é exibida nas seções de parágrafo de um registro. Inclui o título, declaração de responsabilidade, edição, detalhes específicos materiais, informação de publicação, descrição física, série, notas, e números padrões. Os bibliotecários, normalmente, adotam as regras da *Anglo-American Cataloguing Rules* (AACR) [7], que é um padrão para catalogação de documentos.

2. **Entrada principal e entradas adicionais**: A AACR, versão 2, também contém regras para determinar "pontos de acesso" ao registro (normalmente chamado de "entrada principal" e outras "entradas adicionais"), e a forma que estes pontos de acesso deveriam levar. Pontos de acesso são os pontos de recuperação no catálogo da biblioteca, onde os bibliotecários deveriam poder observar o artigo. Em outras palavras, as regras em AACR são usadas para responder perguntas como: Para este livro, deveria haver entradas no catálogo para mais de um autor ou mais de um título? O título da série deveria ser uma nota? Como o nome do autor deveria ser escrito?

```
GV943          Brenner, Richard J., 1941-
.25                        Make the team. Soccer : a heads up guide to super
.B74                  soccer! / Richard J. Brenner. -- 1st ed. -- Boston :
1990                  Little, Brown, c1990.

                                 127 p. : ill. ; 19 cm.

                                 "A Sports illustrated for kids book."
                              Summary: Instructions for improving soccer skills. Discusses
                            dribbling, heading, playmaking, defense, conditioning, mental
                             attitude, how to handle problems with coaches, parents, and
                            other players, and the history of soccer.

                                 ISBN 0316107514 : \$12.95

                                 1. Soccer -- Juvenile literature. 2. Soccer.
                                         II. Title: Heads up guide to super soccer.
                                         II. Title.

                            Dewey Class no.: 796.334/2 -- dc 20         89-48230
                                                                           MARC
```

Figura 1: Exemplo de ficha catalográfica.[17]

3. **Títulos por assunto** (assunto adicionado às entradas): O bibliotecário usa uma lista de eliminação de títulos por assunto, ou alguma outra lista de títulos por assunto padrão para selecionar os assuntos sob os quais o artigo será listado. O uso de uma lista já aprovada e utilizada é importante para garantir consistência e assegurar que todos os artigos são encontrados em um dado assunto sob um mesmo título e, portanto, no mesmo lugar no catálogo. Por exemplo, a lista de título por assunto indica todos os livros sobre gatos associados ao assunto GATOS. Utilizando-se este título elimina-se a possibilidade de listar alguns livros sob o assunto GATOS e outros sob FELINOS. Se um livro é chamado "Tudo Sobre Felinos", o título do assunto será digitado GATOS. Deste modo, serão listados todos os livros naquele assunto em um único lugar no catálogo para o cliente achar. O cliente não tem que imaginar todos os possíveis sinônimos para a palavra ele está procurando.

4. **Número de chamada**: O propósito do número de chamada é colocar artigos juntos no mesmo assunto na mesma estante na biblioteca. A maioria dos artigos é organizada alfabeticamente por autor. A segunda parte do número de chamada normalmente representa o nome do autor.

Na Figura 2, é apresentado um exemplo de um registro MARC, correspondente à ficha da Figura 1.

A Figura 3 mostra o conteúdo de um bloco de um arquivo MARC. Pode-se bservar que as *tags* (destacadas em negrito) não aparecem antes dos campos, mas em um diretório onde é especificada a posição inicial e o tamanho dos campos de dados propriamente ditos.

A primeira *tag*, 001, significa *Control No.*, a segunda, 003, significa *Control No. ID*, a terceira, 005, *DTLT*, e assim por diante (padrão MARC). Os 4 primeiros dígitos que vêm após o trecho em negrito representam o tamanho do campo; os cinco dígitos seguintes representam o ponto inicial para o campo dentro da cadeia de dados que segue o diretório. O caracter $\wedge$ indica um terminador de campo que marca o fim do diretório e dos campos individuais. Os espaços devem ser contados como caracteres na contagem das posições.

TITLE:	Make the team. Soccer : a heads up guide to super soccer! / Richard J. Brenner.
ADDED TITLE:	Heads up guide to super soccer
AUTHOR:	Brenner, Richard J., 1941-
PUBLISHED:	1st ed. Boston : Little, Brown, c1990.
MATERIAL:	127 p. : ill. ; 19 cm.
NOTE:	"A Sports illustrated for kids book."
NOTE:	Instructions for improving soccer skills. Discusses dribbling, heading, playmaking, defense, conditioning, mental attitude, how to handle problems with coaches, parents, and other players, and the history of soccer.
SUBJECT :	Soccer–Juvenile literature. Soccer.
Copies Available:	GV943.25 .B74 1990

Figura 2: Exemplo de registro MARC [17]

```
01041cam  2200265 a 4500001002000000003000400020005001700024008004100041010002400082020002500106020004400131040001800175050002400193082001800217100003200235245008700267246003600354250001200390260003700402300002900439500004200468520022000510650003300730650001200763^###89048230#/AC/r91^DLC^19911106082810.9^891101s1990####maua###j######000#0#eng##^###$a###89048230#/AC/r91^##$a0316107514 :$c$12.95^##$a0316107506 (pbk.) :$c$5.95 ($6.95 Can.)^##$aDLC$cDLC$dDLC^00$aGV943.25$b.B74 1990^00$a796.334/2$220^10$aBrenner, Richard J.,$d1941-^10$aMake the team. $pSoccer :$ba heads up guide to super soccer! /$cRichard J. Brenner.^30$aHeads up guide to super soccer.^##$a1st ed.^##$aBoston :$bLittle, Brown,$cc1990.^##$a127 p. :$bill. ;$c19 cm.^##$a"A Sports illustrated for kids book."^##$aInstructions for improving soccer skills. Discusses dribbling, heading, playmaking, defense, conditioning, mental attitude, how to handle problems with coaches, parents, and other players, and the history of soccer.^#0$aS
```

Figura 3: Bloco de um arquivo no padrão MARC [17]

Depois de separados os campos do bloco, um software apropriado pode exibir um registro da forma como se pode ver na figura 4. Os nomes dos campos, como *Leader*, *Control No.*, *DTLT* e os demais não são armazenados. Eles são exibidos pelo software como uma referência às *tags* que aparecem no arquivo do registro. Por exemplo, na Figura 4, aparece o campo *Control No.*, que é seguido pelo número 001, o qual está armazenado (trecho em negrito) no bloco da Figura 3.

Assim, do diretório podem ser extraídas as informações listadas na Tabela 1:

Tabela 1: Conteúdo do diretório [17]

Tag	Comprimento	Início		Tag	Comprimento	Início
001	0020	00000		100	0032	00235
003	0004	00020		245	0087	00267
005	0017	00024		246	0036	00354
008	0041	00041		250	0012	00390
010	0024	00082		260	0037	00402
020	0025	00106		300	0029	00439
020	0044	00131		500	0042	00468
040	0018	00175		520	0220	00510
050	0024	00193		650	0033	00730
082	0018	00217		650	0012	00763

Pode-se perceber que o bloco da Figura 3 inicia-se com o trecho $01041cam\ \ 2200265\ a\ 4500$, o qual coincide, na Figura 4, com o campo chamado *Leader*. A seguir se inicia o diretório, contendo as *tags* (em negrito)

Do longo trecho numérico após o *Leader*, destaca-se a seqüência $00100200000000300040002000500170002 4$. Pode-se dividir esta seqüência como se segue: $001\ 0020\ 00000\ 003\ 0004\ 00020\ 005\ 0017\ 00024$.

Percebe-se 9 grupos numéricos. Dentro de um cabeçalho de um registro, no formato MARC, estes grupos têm o seguinte significado:

- 001 - significa que os dados a seguir tratam do *Control No.*
- 0020 - é a quantidade de posições para armazenar o valor de *Control No.*
- 00000 - é a posição onde inicia-se o valor de *Control No.*
- 003 - significa que os dados a seguir tratam do Control *No. ID.*
- 0004 - é a quantidade de posições para armazenar o valor de *Control No. ID.*
- 00020 - é a posição inicial do valor *Control No. ID.*
- 005 - significa que os dados a seguir tratam do *DTLT.*
- 0017 - este é a quantidade de posições que o valor do*DTLT* possui.
- 00024 - é a posição inicial do valor do *DTLT.*

Estes parâmetros são apenas para controlar os dados propriamente ditos. Observe-se que, na Figura 3, na 6^a. linha (15^a. posição), encontra-se o caracter $\wedge$. Este símbolo indica o início dos valores que são regidos pelo cabeçalho acima descrito.

```
Leader            01041cam  2200265 a 4500
Control No.       001 ###89048230
Control No. ID    003 DLC
DTLT              005 19911106082810.9
Fixed Data        008 891101s1990    maua  j      001 0 eng
LCCN              010 ## $a ###89048230
ISBN              020 ## $a 0316107514 :
                         $c $12.95
ISBN              020 ## $a 0316107506 (pbk.) :
                         $c $5.95 ($6.95 Can.)
Cat. Source       040 ## $a DLC
                         $c DLC
                         $d DLC
LC Call No.       050 00 $a GV943.25
                         $b .B74 1990
Dewey No.         082 00 $a 796.334/2
                         $2 20
ME:Pers Name      100 1# $a Brenner, Richard J.,
                         $d 1941-
Title             245 10 $a Make the team.
                         $p Soccer :
                         $b a heads up guide to super soccer!
                         $c Richard J. Brenner.
Variant Title     246 30 $a Heads up guide to super soccer
Edition           250 ## $a 1st ed.
Publication       260 ## $a Boston :
                         $b Little, Brown,
                         $c c1990.
Phys Desc         300 ## $a 127 p. :
                         $b ill. ;
                         $c 19 cm.
Note: General     500 ## $a "A Sports illustrated for kids
                            book."
Note: Summary     520 ## $a Instructions for improving soccer
                            skills. Discusses dribbling,
                             heading,
                            playmaking, defense,
                             conditioning,
                            mental attitude, how to handle
                            problems with coaches, parents,
                            and other players, and the
                             history
                            of soccer.
Subj: Topical     650 #0 $a Soccer
                         $v Juvenile literature.
Subj: Topical     650 #1 $a Soccer.
```

Figura 4: Uma melhor visualização de um registro MARC [17]

Considere-se novamente aqueles grupos numéricos citados anteriormente. Os três primeiros indicam que o *Control No.* tem 20 caracteres e se inicia na posição 0. Assim, os 20 caracteres subseqüentes ao sinal $\wedge$ representam valor do *Control No.*, ou seja, ele vale $\sharp\sharp\sharp 89048230\sharp/AC/r91$.

Logo em seguida tem-se o trecho $\wedge DLC$, que corresponde ao quarto, quinto e sexto grupos que foi separado inicialmente, ou seja, o campo *Control No. ID*, que possui 4 caracteres e inicia-se na posição 20 do bloco.

Por fim, aparece o campo *DTLT*, cujo valor 19911106082810.9, pode ser obtido a partir dos três últimos campos da seqüência destacada: 005 0017 00024.

Pode-se, então, construir as três linhas iniciais do registro MARC, mostrado na Figura 4:

```
Control No.       001 ###89048230
Control No. ID    003 DLC
DTLT              005 19911106082810.9
```

Assim, observa-se que o bloco da Figura 3 possui uma primeira parte que representa os metadados do registro, ou seja, os parâmetros que regulamentam a segunda parte do bloco, que representa os dados propriamente ditos.

O MARC é utilizado para catalogações bibliográficas e, aparentemente, não tem relação com a *Web Semântica*. Mas, seu conceito de catalogação é direcionado ao *machine-readable*, isto é, registros para leitura de computadores. Além disso, também utiliza a "definição" de metadados.

A seguir, é apresentado um padrão de metadados direcionado aos meios eletrônicos: o Dublin Core.

5.2 Dublin Core

O padrão Dublin Core (DC) de metadados é um simples, mas efetivo conjunto de elementos para descrever uma ampla quantidade de recursos eletrônicos. O DC compreende quinze elementos semânticos que foram estabelecidos através do consenso de grupos interdisciplinares internacionais de bibliotecários, cientistas da computação, comunidade de museus, e outros estudiosos deste campo.

Outra forma de ver o DC é como uma pequena linguagem para construir classes particulares de declarações sobre os recursos. Nesta linguagem, há duas classes de termos: elementos (nomes) e qualificadores (adjetivos), que podem ser arranjados como um padrão simples de instruções.

No DC, cada elemento é opcional e pode ser repetido. Também tem um número limitado de qualificadores, atributos que podem ser usados em um refinamento posterior para entendimento do elemento que se está qualificando. Um qualificador é qualquer coisa que descreva ou caracterize um objeto. No caso do DC, um qualificador refina o significado do elemento.

O DC tem como objetivo seguir as seguintes características:

- Simplicidade de criação e manutenção: o conjunto de elementos foi mantido pequeno a fim de permitir que um não-especialista crie registros descritivos simples facilmente, enquanto fornece recursos de recuperação efetiva no ambiente conectado em rede.

- Semântica comumente compreendida: a descoberta de informações na Internet é dificultada pelas diferenças de terminologias de um campo de conhecimento para outro. O

usuário poderá encontrar o que procura através da capacidade dos elementos comuns de semântica, que são universalmente compreendidos e suportados pelo DC.

- Escopo internacional: O conjunto de elementos do DC foi originalmente desenvolvido em inglês, mas versões foram sendo criadas em outras linguagens, incluindo finlandês, norueguês, tailandês, japonês, francês, português, alemão, grego, indonésio, e espanhol. Embora os desafios técnicos de internacionalização sobre a WWW não sejam diretamente endereçados para a comunidade de desenvolvimento do DC, o envolvimento de representantes de quase todos os continentes assegurou que o desenvolvimento do padrão fosse considerado de natureza multilíngüe e multicultural no universo de informação eletrônica.

- Extensibilidade: apesar dos esforços de manter a simplicidade na descrição dos recursos digitais, o DC reconhece a importância de fornecer mecanismos de extensão para necessidade de recursos adicionais descobertos. Espera-se que outras comunidades de metadados criem e administrem conjuntos adicionais de metadados. Estes elementos poderiam ligar-se ao conjunto do DC para satisfazer esta necessidade de extensão. Desta forma, seria possível que diferentes comunidades utilizassem os elementos do DC para especificar informações descritivas que fossem úteis através da Internet.

É apresentada, a seguir, uma breve descrição dos quinze elementos do DC, de acordo com [20]:

1. *Title* (Título) - O nome dado ao documento eletrônico pelo autor ou editor.

2. *Author or Creator* (Autor) - Pessoas ou organizações responsáveis pelo conteúdo intelectual do objeto. (Ex.: autores no caso de documentos escritos; artistas, fotógrafos ou ilustrador no caso de recursos visuais).

3. *Subject and Keywords* (Assunto) - Representa o assunto do documento eletrônico, podendo ser definido a partir de sistemas de classificação (CDD − Classificação Decimal de Dewey, CDU − Classificação Decimal Universal, LCSH − *Library of Congress Subject Headings*) ou simplesmente por uma palavra ou conjunto de palavras.

4. *Description* (Descrição) - Descrição do conteúdo, podendo ser resumo ou descrição no caso de recursos visuais.

5. *Publisher* (Editor) - Entidades responsáveis por tornar o documento disponível na presente forma, tais como editor, universidades ou entidades corporativas.

6. *Other Contributors* (Outros Colaboradores) - Outras pessoas que contribuíram para a realização da obra (editores, tradutores, ilustrador etc.).

7. *Date* (Data) - A data em que o documento foi disponibilizado na presente forma.

8. *Resource Type* (Tipo de recurso) - Gênero do recurso, tais como: *home page*, novela, poema, dicionário, software aplicativo, arquivo de dados etc.

9. *Format* (Formato) - A manifestação física do documento eletrônico, tal como: Postscript, PDF ou HTML.

10. *Resource Identifier* (Identificação) - Série ou número usado para identificar o documento (URL, ISBN etc.).

```
...
<head profile="http://dublincore.org/documents/dcq-HTML/">
<title>Expressing Dublin Core in HTML/XHTML meta and link
    elements</title>
<link rel="schema.DC" href="http://purl.org/dc/elements/1.1/" />
<link rel="schema.DCTERMS" href="http://purl.org/dc/terms/" />

<meta name="DC.title" lang="en" content="Expressing Dublin Core
         in HTML/XHTML meta and link elements" />
<meta name="DC.creator" content="Andy Powell, UKOLN,
   University of Bath" />
<meta name="DCTERMS.issued" scheme="DCTERMS.W3CDTF"
         content="2003-11-01" />
<meta name="DC.identifier" scheme="DCTERMS.URI"
         content="http://dublincore.org/documents/dcq-HTML/" />
<link rel="DCTERMS.replaces" hreflang="en"
href="http://dublincore.org/documents/2000/08/15/dcq-HTML/" />
<meta name="DCTERMS.abstract" content="This document describes how
        qualified Dublin Core metadata can be encoded
        in HTML/XHTML &lt;meta&gt; elements" />
<meta name="DC.format" scheme="DCTERMS.IMT" content="text/HTML" />
<meta name="DC.type" scheme="DCTERMS.DCMIType" content="Text" />
</head>
...
```

Figura 5: Registro Dublin Core embutido em HTML [19]

11. *Source* (Fonte) - O documento (impresso ou eletrônico) do qual se originou o recurso eletrônico.

12. *Language* (Idioma) - Idioma do conteúdo intelectual do documento.

13. *Relation* (Relação) - Relacionamento com outros documentos impressos ou eletrônicos (por exemplo imagens em um documento, capítulos em um livro ou itens em uma coleção).

14. *Coverage* (Cobertura) - Locação espacial ou duração temporal característica do documento.

15. *Rights Management* (Direito Autoral) - Informação sobre *copyright*.

Pode ser visto, na Figura 5, um exemplo de um registro Dublin Core ou DC. Observa-se que ele está dentro de um código HTML:

Um registro DC pode ser representado também em RDF (*Resource Description Framework*). RDF [15] é uma tecnologia desenvolvida pelo W3C destinada a prover meio de intercâmbio de metadados. Constitui-se em uma arquitetura genérica de metadados que permite descrever recursos no contexto Web, através da adoção de padrões de metadados. RDF não é uma linguagem, mas um modelo de dados para descrição de recursos com mais semântica, através da adoção de metadados. Utiliza o conceito de sentença. Uma sentença é um par, propriedade-valor e um recurso ao qual essa propriedade se aplica.

```
1  <rdf:RDF
2     xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
3     xmlns:dc="http://purl.org/dc/elements/1.1/">
4
5     <rdf:Description
6        rdf:about="http://media.example.com/audio/guide.ra">
7        <dc:creator>Rose Bush</dc:creator>
8        <dc:title>A Guide to Growing Roses</dc:title>
9        <dc:description>
10          Describes process for planting and nurturing
11          different kinds of rose bushes.
12       </dc:description>
13       <dc:date>2001-01-20</dc:date>
14
15    </rdf:Description>
16 </rdf:RDF>
```

Figura 6: Registro Dublin Core Expresso em XML/RDF [13]

Na Figura 6, nas linhas 2 e 3, faz-se a declaração de quais fontes de metadados (*namespaces*) o registro está incorporando. Na linha 2 (xmlns:rdf) o registro incorpora a sintaxe referente ao endereço indicado (“http://www.w3.org/1999/02/22−rdf−syntax−ns#”). O mesmo vale para a linha 3 (xmlns:dc) que referencia o endereço “http://purl.org/dc/elements/1.1/” para utilizar as propriedades do padrão Dublin Core.

6 Tipos de Metadados

Metadados podem possuir a classificação do tipo estrutural ou semântico. Metadado estrutural representa a informação que descreve a organização e estrutura dos dados gravados; por exemplo, informações sobre o formato, os tipos de dados usados e os relacionamentos sintáticos entre eles.

Em contraste, metadados semânticos fornecem informações sobre o significado dos dados disponíveis e seus relacionamentos semânticos; por exemplo, dados que descrevem o conteúdo semântico de um valor de dado (como unidades de medida e escala), ou dados que fornecem informações adicionais sobre sua criação (algoritmo de cálculo ou derivação da fórmula usada), linhagem dos dados (fontes) e qualidade (atualidade e precisão). Faz-se necessário elaborar um modelo de metadados que descreva o contexto da informação de uma maneira não ambígua ou redundante.

Uma conceitualização de um domínio específico de problema ou ontologias [12] que forneçam um acordo comum de vocabulários para que os dados sejam referenciados é desejável. Assim, uma ontologia serve como uma base comum para a representação de dados e metadados.

As fontes de dados descrevem informações equivalentes diferentemente. Elas fornecem diferentes aspectos da mesma informação, e representam o mesmo aspecto do mundo real, usando diferentes construções estruturais ou conceitos semânticos. Um objeto semântico representa um item de dado junto com sua base de contexto semântico que consiste de um conjunto flexível de meta-atributos que explicitamente descrevem a compreensão implícita sobre o significado do item de dado.

Adicionalmente, cada objeto semântico possui um rótulo de conceito associado a ele, que especifica o relacionamento entre o objeto e os aspectos do mundo real que ele descreve. Estes rótulos são adquiridos de uma ontologia. A detecção e resolução destas heterogeneidades semânticas, obviamente, requerem conhecimento sobre a exata base semântica dos dados representados.

Para assegurar a correta interpretação dos metadados disponíveis um domínio específico de ontologias pode ser utilizado. Uma ontologia fornece compreensão sobre uma conceitualização compartilhada de um determinado domínio de aplicação. Os conceitos específicos numa ontologia fornecem vocabulário comum para que nenhuma negociação adicional seja necessária. Além disso, a ontologia fornece informação sobre a representação do dado descrito sobre a base do modelo. Numa situação ideal, todas as instâncias que fazem uso dos dados e metadados de um determinado domínio devem aderir à ontologia correspondente.

7 Considerações Finais

Foram apresentadas neste trabalho algumas considerações sobre metadados, suas aplicações e seus padrões. O que se pode perceber é a grande utilidade que há nas diversas formas de aplicação e utilização, seja na *web*, seja em sistemas proprietários ou legados. Daí a confirmação que metadados ainda estão em ascensão e que novos estudos são necessários.

As tecnologias XML e RDF, assim como padrões como Dublin Core e MARC já têm grande abrangência tanto na academia como no mercado, demonstrando sua grande utilidade.

Os metadados irão adquirir muito mais importância com a evolução das tecnologias associadas à *Web Semântica*, tornando-se um componente crítico para qualquer arquitetura.

8 Agradecimento

A Profa. Dra. Ana Paula Laboissière Ambrósio, pela avaliação do presente texto e pelas sugestões feitas, as quais muito contribuiram para a melhoria do texto original.

Referências

[1] **Google**. http://www.google.com.br, acessado em outubro de 2004, 2004.

[2] CIMI. **Consortium for the Interchange of Museum Information**. http://www.cimi.org/, acessado em outubro de 2004, 2004.

[3] CODD., E. F. **The relational model for database management**. Addson-Wesley Publishing Company, 1990.

[4] COMITEE, F. G. D. **Metadata**. http://www.fgdc.gov/metadata/metadata.html, acessado em outubro 2004, 2004.

[5] CORE, D. **Dublin Core Metadata Initiative**. http://www.dublincore.org, acessado em outubro de 2004, 2004.

[6] CROSS, P; RAHIMI, S. **The Open Information Model**. SQL Server Magazine, March 2000.

[7] FOR REVISION OF ANGLO-AMERICAN CATALOGUING RULES (AACR), J. S. C. **Documents**. http://www.collectionscanada.ca/jsc/docs.html#logical, acessado em outubro de 2004, 2004.

[8] GCMD, G. C. M. D. **Directory Interchange Format (DIF). Writer's Guide, Version 9**. http://gcmd.gsfc.nasa.gov/User/difguide/difman.html , acessado em outubro de 2004, 2004.

[9] GILS. **Government Information Locator Service (GILS).** http://www.access.gpo.gov/su_docs/gils/, acessado em outubro 2004, 2004.

[10] GROUP, O. M. **Common Warehouse MetaModel** . http://www.omg.org/cwm/, acessado em outubro de 2004, 2004.

[11] GROUP, O. M. **XML Metadata Interchange (XMI)** . http://www.omg.org/technology/documents/formal/xmi.htm, acessado em outubro de 2004, 2004.

[12] GRUNINGER, M; LEE, J. **Ontology: Applications and Design**. Communications of the ACM, 45(2), 2002.

[13] HILLMANN, D. **Using Dublin Core**. http://dublincore.org/documents/2003/08/26/usageguide/, acessado em outubro de 2004, 2004.

[14] IKEMATU, R. S. **Gestão de metadados: sua evolução na tecnologia da informação**. Data Grama Zero - Revista de Ciência da Informação, 2(6), 2001.

[15] MANOLA, F; MILLER, E. E. **RDF Primer. W3C Recommendation 10 February 2004.** http://www.w3.org/TR/rdf-primer/, acessado em outubro de 2004, 2004.

[16] OF APPLICATION SPECIFIC SIGNAL PROCESSORS (RASSP), R. P. **Metadata Interchange Specification. Versão 1.1 de 01/08/1997** . http://www.eda.org/rassp/documents/atl/MDIS-11.pdf, acessado em outubro de 2004, 2004.

[17] OF CONGRESS, L. **MARC 21 Reference Materials**. http://www.loc.gov/marc/umb/um11to12.html, acessado em outubro de 2004, 2004.

[18] OF CONGRESS, L. **MARC Standards**. http://www.loc.gov/marc/, acessado em outubro de 2004, 2004.

[19] POWELL, A. **Expressing Dublin Core in HTML/XHTML meta and link elements**. http://www.dublincore.org/documents/dcq-html/, acessado em outubro de 2004, 2004.

[20] ROSETTO, M; NOGUEIRA, ADRIANA, H. **Aplicações de elementos de metadados Dublin Core para descrição de dados bibliográficos *on-line* da biblioteca digital de teses da USP**. http://www.sibi.ufrj.br/snbu/snbu2002/oralpdf/82.a.pdf, acessado em junho de 2004, 2004.

[21] W3C. **Extensible Markup Language (XML) 1.0 (Third Edition). W3C Recommendation 04 February 2004**. http://www.w3.org/TR/REC-xml/, acessado em outubro de 2004, 2004.

[22] W3C. **Extensible Stylesheet Language (XSL)**. http://www.w3.org/Style/XSL/, acessado em outubro de 2004, 2004.

[23] W3C. **Home Page**. http://www.w3.org/ , acessado em outubro de 2004, 2004.

[24] W3C. **Semantic Web**. http://www.w3.org/2001/sw/ , acessado em outubro de 2004, 2004.

[25] W3C. **XML in 10 Points**. http://www.w3.org/XML/1999/XML-in-10-points.html, acessado em outubro de 2004, 2004.