UNIVERSIDADE FEDERAL DE RIO DE JANEIRO
Escola de Comunicação
INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA E TECNOLOGIA
Programa de Pós-graduação em Ciência da Informação

METADADOS PARA RECUPERAÇÃO DA INFORMAÇÃO

EM AMBIENTE VIRTUAL

Tese apresentada como requisito parcial para obtenção do título de Mestre em Ciência da Informação

Autora: MÔNICA CRISTINA COSTA SANTIAGO

Orientadora: Profª Lena Vania Ribeiro Pinheiro
Drª em Comunicação e Cultura, UFRJ/ECO

Rio de Janeiro
2004

MÔNICA CRISTINA COSTA SANTIAGO

METADADOS PARA RECUPERAÇÃO DA INFORMAÇÃO EM AMBIENTE VIRTUAL

Tese apresentada como requisito parcial para obtenção do título de Mestre em Ciência da Informação, ao Programa de Pós-Graduação em Ciência da Informação da Universidade Federal do Rio de Janeiro –UFRJ/Escola de Comunicação, em convênio com o Instituto Brasileiro de Informação em Ciência e Tecnologia-IBICT.

Rio de Janeiro
2004

025.04 Santiago, Mônica Cristina Costa.
S235 Metadados para recuperação da informação em ambiente virtual / Mônica Cristina Costa Santiago. – Rio de Janeiro, 2004.
ix, 111 f. : il.

Dissertação (mestrado em Ciência da Informação). UFRJ/ECO-MCT/IBICT. Orientadora: Lena Vania Ribeiro Pinheiro.

1. Recuperação da informação. 2. Internet – Programas de Computador. 3. Metadados. I. Pinheiro, Lena Vania Ribeiro. II. Título.

METADADOS PARA RECUPERAÇÃO DA INFORMAÇÃO EM AMBIENTE VIRTUAL

Mônica Cristina Costa Santiago

Dissertação submetida como parte dos requisitos para obtenção do título de Mestre em Ciência da Informação.

Aprovada por:

_____

Profª. Hagar Espanha Gomes
Livre docente, UFF.

_____

Profª. Maria de Nazaré Freitas Pereira
Doutora em Ciências Humanas, IUPERJ

_____

Profª. Lena Vania Ribeiro Pinheiro – Orientadora
Doutora em Comunicação e Cultura, UFRJ/ECO

SUPLENTE:

_____

Profª. Rosali Fernandez de Sousa (IBICT)
Ph.D. In Information Science (Polytechnic of North London, England)

AOS MEUS AVÓS AIDA, VIOLANDA, LUIZ,

ROSA LÍDIA E HERCULANO

COM SAUDADES.

AGRADECIMENTOS

Ao meu marido Walter, pela sua solidaridade irrestrita e amor, sem os quais não teria sido possível realizar o trabalho.

Aos meus pais Hélio e Sônia, que sempre me apoiaram e me compreenderam em todos os momentos.

Aos meus tios Solange, Vilma e Sérgio, que me acompanharam em todas as fases de minha vida, sempre demostrando confiança e apoio em minhas escolhas.

Aos meus sogros Rosa e Froilán, pela compreensão e carinho com que me acolheram.

À minha querida amiga Selma Chi Barreiro, que me incentivou a cursar o mestrado e me apoiou incondicionalmente ao longo do curso.

À minha querida orientadora Professora Lena Vania Ribeiro Pinheiro, que me compreendeu quando do meu afastamento e que me acolheu de braços abertos no meu retorno ao Brasil, encorajando-me a retomar meus estudos, e pelo seu apoio irrestrito na revisão e conclusão do trabalho.

À minha colega de turma Carla Tavares, pelo carinho e amizade.

À empresa Documentar, representada pela Gerente de Projetos Especiais, Marília Rocha, que entendeu e possibilitou meu afastamento para a finalização da dissertação.

Por último, mas não menos importante, agradeço a Deus, que me concedeu a grande felicidade de ter concluído mais esta etapa da minha vida.

RESUMO

SANTIAGO, Mônica Cristina Costa. Análise de metadados para recuperação da informação em ambiente virtual. Rio de Janeiro, Universidade Federal de Rio de Janeiro, Escola de Comunicação-UFRJ-ECO; Instituto Brasileiro de Informação em Ciência e Tecnologia-IBICT, 2004. Dissertação (Mestrado em Ciência da Informação). Orientadora: Lena Vania Ribeiro Pinheiro.

Análise de metadados no exterior e no Brasil, tendo a catalogação, classificação e indexação como fundamentos teóricos e técnicos, nas suas inter-relações, e com foco nos sistemas de recuperação da informação, acompanhados na sua trajetória evolutiva desde sistemas manuais, automatizados até a *Internet/Web*. No ambiente virtual a recuperação da informação é estudada, com seus critérios de avaliação e instrumentos como esquemas de classificação bibliográfica, cabeçalhos de assuntos e tesauros. Os metadados são abordados nos seus conceitos, definições, tipos, características e funções e, nos esquemas identificados, a sintaxe e interoperabilidade são destacadas. Os resultados referem-se ao mapeamento de metadados no Brasil e em outros países, com ênfase no Dublin Core. As conclusões apontam o conhecimento de metadados e seu uso no Brasil, sobretudo o MARC e Dublin Core.

ABSTRACT

SANTIAGO, Mônica Cristina Costa. Análise de metadados para recuperação da informação em ambiente virtual. Rio de Janeiro, Universidade Federal de Rio de Janeiro, Escola de Comunicação-UFRJ-ECO; Instituto Brasileiro de Informação em Ciência e Tecnologia-IBICT, 2004. Dissertação (Mestrado em Ciência da Informação). Orientadora: Lena Vania Ribeiro Pinheiro.

This research analyses metadata use in Brazil and abroad, based on cataloging, classification and indexing theory and techniques, focused on information retrieval system and its evolution, from manual, automated systems till internet/web. In the virtual environment, the information retrieval, its evaluation criteria and tools like classification schemes, subject headings and thesaurus, are studied. Metadata concepts, definitions, types and attributes are presented and syntax and interoperability are the focal point in the identified metadata schemes. The results refer to metadata use mapping in Brazil and abroad, stressing the Dublin Core. The conclusion points out metadata knowledge and use in Brazil, specially Mark and Dublin Core.

SUMARIO

LISTA DE FIGURAS, QUADROS E TABELAS

1 INTRODUÇÃO

O mundo vem sofrendo mudanças intensas nas últimas décadas do século XX, produzidas e difundidas velozmente em todo o globo do ponto de vista social, econômico, cultural e político. Esta nova ordem, caracterizada por uma série de grandes transformações, vem sendo denominada de inúmeras formas, quais sejam: Economia ou Sociedade Informacional, Novo Regime de Acumulação e Regulação, Paradigma Tecno-Econômico das Tecnologias de Informação e Comunicação ou, de acordo com Lastres e Ferraz (1999), Era, Economia ou Sociedade do Conhecimento ou do Aprendizado.

Para Castells (1999, p. 50): "O cerne da transformação que estamos vivendo na revolução atual refere-se às tecnologias de informação, processamento e comunicação".

Nestas transformações, as tecnologias de informação e comunicação são fundamentais. O termo Tecnologias da Informação se refere a diferentes áreas, entre outras, à Informática, Telecomunicações, Ciência da Computação, Ciência da Informação, Engenharia de Sistemas e de *Software* e, segundo Lastres e Ferraz (1999, p. 33):

> O novo paradigma das tecnologias de informação é visto como baseado em um conjunto interligado de inovações em computação eletrônica, engenharia de *software*, sistemas de controle, circuitos integrados e telecomunicações, que reduziram drasticamente os custos de armazenamento, processamento, comunicação e disseminação da informação.

O imperativo tecnológico é responsável por gerar e impulsionar o desenvolvimento e aplicação de um grande número de serviços de informação, produtos, sistemas e redes, mas a base das transformações atuais, segundo Saracevic, é a importância dos papéis desempenhados pela informação e pelo conhecimento na sociedade globalizada. Drucker (apud SARACEVIC, 1995, p. 36) demonstra a extensão destas mudanças que estão desafiando a tradicional teoria econômica de valor:

> O recurso econômico básico – "os meios de produção", para empregar o termo utilizado pelos economistas – não é mais o capital, ou os recursos naturais (a "terra" dos economistas), nem o "trabalho". É e será o conhecimento ... O valor agora é gerado pela "produtividade" e "inovação", ambas aplicações do conhecimento para o "trabalho". O grupo social dominante da sociedade da informação será representado pelos "trabalhadores do conhecimento" – executivos do conhecimento que sabem como alocar conhecimento para uso produtivo, da mesma forma que os capitalistas sabiam como alocar capital para este fim; profissionais do conhecimento, funcionários do conhecimento ... O desafio econômico da sociedade pós-capitalista será, portanto, a produtividade do trabalho para o conhecimento e do trabalhador do conhecimento.

No entanto, as transformações ocorrem em todos os setores, atingindo a sociedade como um todo. Para entendimento dessas tecnologias e principalmente do novo "espaço",

caracterizado exatamente pela desterritorialização, alguns teóricos muito o têm estudado e publicado obras a respeito, entre os quais Michel Serres e Pierre Lévy, o último muito adotado no Brasil, país que tem visitado algumas vezes, para aulas e conferências.

Lévy (1996) assim se expressa sobre o fenômeno e sua relação com registros, comunicação e rapidez:

> De maneira análoga, diversos sistemas de registro e de transmissão (tradição oral, escrita, registro audiovisual, redes digitais) constroem ritmos, velocidades ou qualidades de história diferentes... Cada novo 'agenciamento', cada máquina tecnossocial acrescenta um espaço-tempo, uma cartografia especial, uma música singular...

Assim, com o desenvolvimento acelerado dessas tecnologias de informação, o encurtamento das distâncias e a velocidade e interatividade da *Internet*, o processo de disseminação da informação ficou bastante facilitado e, no Brasil, alguns autores também se manifestam sobre a questão. Segundo Sayão (2000, p. 146), a nova era tem como característica "o aumento extraordinário da capacidade humana de ampliar seus conhecimentos, de armazená-los, transformá-los, organizá-los e difundí-los instantaneamente".

Todos estes avanços foram possíveis principalmente a partir da *Internet*, que causou uma revolução em termos de disponiblidade e rapidez ao acesso à informação, trazendo em si a superação das fronteiras espaço-temporais, ao promover interações independentemente dos limites físicos e a interconexação entre diferentes redes de computadores, permitindo a qualquer interessado o acesso, diretamente de seu computador pessoal, da informação de que se necessita (SAYÃO, 2000).

O advento da *Web*, a grande rede mundial de computadores, causou uma verdadeira revolução no mundo da recuperação da informação, trazendo à tona novas metodologias e abordagens. Atualmente, a *Internet* é utilizada em todas as esferas da vida para a troca de informações. As bibliotecas agora oferecem seus *Online Public Access Catalogues* (OPACs) na *Internet*, os vendedores de bases *online*, tais como *Silver Platter*, tornam suas bases de dados acessíveis através da *Internet*; todos os tipos de organizações (nacionais, internacionais, educacionais, de pesquisa e comerciais), tornam diferentes tipos de informação acessíveis na *Web*. A mudança do ambiente tem significativas implicações para o mundo da recuperação da informação como um todo, e como conseqüência, profissionais da informação enfrentam novos desafios. Além disso, no momento, estamos experimentando uma explosão da disponibilidade de informação eletrônica com a proliferação de páginas individuais e institucionais (CHOWDHURY, 1999).

Bellcore (1995, p.10) corrobora as idéias de Chowdhury ao afirmar que: "o que é notável não é o fato de que todos estão acessando informação, mas sim que todos estão disponibilizando informação. Por décadas, a distribuição de informação tinha estado nas mãos de alguns, enviando informações para muitos usuários. Agora, os usuários estão gerando sua própria informação e classificando-a eles próprios, já que todos os tipos de pessoas criam sua própria *homepage* e a *linkam* com todos os tipos de recursos desejados".

Mas, se por um lado há disponibilidade e acesso rápido ao repositório de informações da *Internet*, por outro, navegar em suas páginas e achar o que se quer é considerada uma tarefa de sorte: o volume de informações é muito grande, conseqüentemente, muito tempo é gasto para encontrar o que se procura. O fenômeno da explosão de documentos eletrônicos, que será mais detalhado no capítulo 4, é atestado pelo estudo de Lyman e Varian (2003) sobre a quantidade de informação disponível na *Web*: a *World Wide Web* contém cerca de 170 *terabytes* de informação em sua superfície; em volume isto é 17 vezes maior que o volume das coleções impressas da *Library of Congress*.

Entretanto, o problema do grande volume de informações não é novo, já era um desafio a ser enfrentado em 1945, retratado por Vannevar Bush, em artigo intitulado "As We May Think" e, para combater os impasses da "explosão de informação", ele apresenta como solução o uso da tecnologia da informação.

Outro autor clássico da Ciência da Informação que abordou também este fenômeno foi Bradford, que em seu livro *Documentation*, de 1948, criou a expressão "caos documentário", ao se referir ao volume maciço de informações.

Além do problema da grande quantidade de informações, os usuários se deparam com outras dificuldades quando navegam a *Web*, algumas delas apontadas por Perez (2000):

- Abundância de informações: o resultado de uma busca oferece um número de documentos desmensurado, impossível de se visualizar.
- Pouca relevância: grande parte dos resultados oferecidos não interessa e isso provoca a perda de tempo e a desilusão do usuário.
- Pouca confiabilidade dos resultados: muitos *links*[1] não funcionam, desconhecimento da qualidade da fonte e muitas vezes da própria autoria.

[1] *Link* é um elo de ligação entre dois elementos que, estando em ambiente eletrônico, emprega recursos hipertextuais ou de hipermídia.

- Escassez de recursos de busca: os usuários não sabem como fazer a busca para obter somente aquilo que mais lhes interessa, ou porque o sistema de recuperação não o permite ou porque os usuários desconhecem.

De nada adianta ter abundância de informações, é necessário dispor de informações relevantes. Neste sentido, o conceito de relevância de Saracevic é um dos mais importantes na Ciência da Informação, como bem o demonstram Pinheiro e Loureiro (1995, p. 45):

> Saracevic distingue informação de informação relevante, esta última relacionada a mecanismos de comunicação seletiva e à orientação aos usuários de sistemas de recuperação da informação. A efetividade da comunicação do conhecimento se dá, segundo Saracevic, na medida de transmissão de um arquivo ao outro, ocasionando mudanças. Portanto, relevância é a medida de tais mudanças, e a Ciência da Informação, ao lado da lógica e da filosofia, apresenta-se como disciplina essencial nos territórios dos estudos e reflexões sobre relevância e, conseqüentemente, informação.

Mas a informação relevante precisa ser encontrada por quem dela necessita, o que nos remonta ao pensamento de Shialy Ramamrita Ranganathan, bibliotecário indiano que elaborou as 05 leis que fundamentam a biblioteconomia: "os livros são para serem usados", "a cada leitor o seu livro", "para cada livro o seu leitor", "poupe o tempo do leitor", "a biblioteca é uma organização em crescimento". Quando enunciadas, as Cinco Leis da Biblioteconomia se restringiam ao contexto da Biblioteca, mas atualmente elas podem ser perfeitamente aplicadas em todos os serviços de informação, que envolvem as atividades de profissionais situados entre o produtor de conhecimento e o usuário da informação (CAMPOS, 2004).

Diante desta realidade, torna-se imprescindível o desenvolvimento de padrões que visem a descrição exata dos recursos de informação em meio eletrônico, pois, segundo Pinheiro (2002, p. 7), "grandes volumes de dados e intercâmbio de informação têm nos padrões a condição *sine qua non* para recuperação e intercâmbio de informações".

Dentre as soluções preconizadas para dar ordem ao *caos* da *Web*, existem os metadados, que podem, genericamente, ser definidos como dados sobre dados. Os metadados criam uma estrutura para a descrição padronizada de documentos, com o objetivo de tornar possível e mais eficiente a identificação, caracterização e localização das informações disponíveis na *Web* (SOUZA, CATARINO e SANTOS, 1997).

Pela urgência do tema e necessidade de uma análise mais aprofundada sobre metadados, a partir do olhar da Ciência da Informação, esta pesquisa foi realizada com o intuito de contribuir para o melhor entendimento das funções dos metadados para a recuperação da informação na *Web*, tanto no Brasil como no exterior, de forma a se constituir em instrumental para os

profissionais da informação no Brasil envolvidos com a criação, manutenção e utilização dos metadados.

A escolha do tema desta pesquisa foi resultado natural da atuação profissional da mestranda como Analista da Informação durante três anos no Programa Prossiga, mais especificamente no Serviço Páginas Brasileiras. O desconhecimento, em geral, sobre metadados e o desejo em adquirir competência e conhecimento acerca do ambiente virtual da *Web* foram fatores que também suscitaram o interesse pelo tema.

Cabe abordar algumas dificuldades da pesquisa, decorrentes da natureza do objeto estudado: a análise mais aprofundada de metadados apresenta um complicador, pois sendo tema relativamente novo, está sofrendo e vai sofrer uma série de modificações, a partir de pesquisas e aplicações.

Outra dificuldade é que metadados constituem uma questão interdisciplinar, não passam apenas pela Ciência da Informação, mas também pela compreensão das tecnologias e campos adjacentes, perpassando, portanto, a Ciência da Computação, entre outros.

Nesta dissertação, muito naturalmente, os metadados são estudados sob a ótica da Ciência da Informação, reconhecida como área "... participante ativa e deliberada da Sociedade da Informação, assim como outras áreas, mas que tem um papel fundamental a exercer, pela sua dimensão social e humana, acima e além da tecnologia" (SARACEVIC, 1992, p. 1). Para tal, utilizamos téoricos importantes da Ciência da Informação como os já citados Saracevic, Borko e Bradford e fazemos referência a precursores da área como Paul Otlet e Vannevar Bush.

Em nossa análise sobre as técnicas de indexação, classificação e catalogação, utilizamos especialistas brasileiros de renome como Barbosa (1978), Campos (2001, 2004), Gomes H. (1997, 2000), Piedade (1983) e Robredo e Cunha (1986). Para o estudo do sistema de recuperação da informação foram escolhidos autores como Lancaster (1979, 1993), um dos teóricos mais importantes, Harter (1986) e Palmer (1987), além de Robredo e Cunha (1986), já mencionados. Sobre metadados, além de autores como Milstead e Feldman (1999), Medeiros (1999), dentre outros, cujos artigos estão disponíveis na rede, recorremos, também, a especialistas cujas obras tratam unicamente sobre o tema e não estão traduzidas para o português, como é o caso de Caplan (2003), Hudgins et al. (1998) e ainda Weber (2002). Foi difícil não incluir, para finalizar o trabalho, informações que descobríamos a todo momento na *Internet*, pelo interesse que o tema vem despertando e seu estado de ebulição em diferentes campos, inclusive na Ciência da Informação.

Esta pesquisa tem os seguintes objetivos:

Objetivo Geral:

Analisar o papel dos metadados nos Sistemas de Recuperação da Informação na *Web*, no contexto de suas transformações a partir das Tecnologias de Informação e Comunicação, sob o olhar da Ciência da Informação.

Objetivos Específicos:

- Analisar os conceitos e definições de metadados e estudar as suas relações e interdependência com a catalogação, a classificação e a indexação tradicional/convencional;
- Levantar e descrever os esquemas de metadados existentes e suas características, incluindo interoperabilidade; e
- Verificar a utilização de metadados em sistemas de recuperação da informação na *Web*, no Brasil.

Esta pesquisa tem caráter teórico-conceitual e empírico. Na etapa teórico-conceitual, foi realizado estudo de definições e conceitos e sua evolução, bem como a interdependência com outros conceitos, contemplando questões relacionadas à temática da pesquisa a aos objetivos estabelecidos, aqui descritos.

A etapa empírica, por sua vez, apresenta duas partes. Na primeira, estudamos o padrão Dublin Core, escolhido por ser um dos primeiros padrões específicos para a descrição de recursos de informação na *Web*. O Dublin Core é um padrão internacionalmente reconhecido, cuja importância pode ser demonstrada pelos estudos desenvolvidos pela própria *Online Computer Library Center* (OCLC)[2], uma das maiores redes prestadoras de serviços de informação nos Estados Unidos, responsável por promover uma rede cooperativa internacional de bibliotecas, de grande importância, mediante a utilização do formato MARC para catalogação bibliográfica. É por esta razão que ao analisarmos o Dublin Core, também fazemos um contraponto com o formato MARC, utilizado pelas bibliotecas por muito tempo, considerado por muitos um padrão mais complexo, tentando identificar vantagens e desvantagens da aplicação de ambos.

Para retratarmos a utilização de metadados no exterior, além do Dublin Core e do MARC, mapeamos outros vários esquemas/padrões, escolhidos por representarem diversas comunidades e terem diferentes aplicações, totalizando 27 esquemas. Este mapeamento é apresentado no Anexo 1, que pode funcionar como um guia para os interessados sobre o

[2] Um dos projetos de pesquisa da OCLC, em conjunto com o *Dublin Core Metadata Initiative Registry* (DCMI), é o desenvolvimento do *Dublin Core Metadata Registry*.

assunto, contendo as seguintes informações: definição/objetivo, instituições responsáveis, comunidades atendidas, *homepage* do esquema e *URLs* para acesso dos elementos relacionados. Estas informações foram coletadas nos *sites* oficiais de cada um dos esquemas, quando identificadas nas *homepages* analisadas.

Na segunda parte da etapa empírica, atendendo ao objetivo específico foi verificada a utilização de metadados em serviços brasileiros de informação na *Web* e seus respectivos sistemas de recuperação da informação. Sobre esta etapa da análise empírica, nossa intenção à época da apresentação do projeto de pesquisa, era verificar, inicialmente, quais serviços de informação eletrônica no Brasil, tais como bibliotecas virtuais/digitais e os *Online Public Acess Catalogues* (OPACs) adotavam padrões de metadados ou até os aplicavam de forma adaptada. No projeto, citamos como exemplos de sistemas que utilizam o padrão Dublin Core no Brasil, a Biblioteca Nacional de Teses e Dissertações da USP (ROSETTO e NOGUEIRA, 2002) e a Embrapa Informática Agropecuária, com o banco de imagens Rural Mídia (SOUZA, VENDRÚSCULO e MELO, 2000).

No caso das Bibliotecas Digitais, o foco seria o Prossiga Informação e Comunicação para a Ciência e Tecnologia. A paralisação e praticamente desativação do Prossiga, pelo menos por enquanto, acarretou uma mudança nas fontes e procedimentos. Desta forma, decidimos analisar o universo dos sistemas de informação de bibliotecas universitárias brasileiras. Para esta etapa da análise empírica, apresentamos a seguir os procedimentos metodológicos adotados:

1. elaboração de um questionário: na primeira parte do questionário, o objetivo é investigar o grau de conhecimento do profissional responsável pelo preenchimento do mesmo a respeito dos metadados, inclusive para identificar quais são os esquemas mais conhecidos por ele. A segunda parte do questionário tem como foco coletar informações referentes à utilização ou não dos metadados, com o intuito de definir qual o padrão utilizado pelos sistemas de informação das bibliotecas. O modelo de questionário enviado encontra-se no Anexo 2;
2. escolha do universo: bibliotecas de universidades federais e eventualmente estaduais que possuem sistemas de informação disponíveis na *Web*. Além deste critério de seleção, utilizamos também a base de dados cadastral da Comissão Brasileira de Bibliotecas Universitárias (CBBU), como fonte formal de informação para a escolha das bibliotecas;
3. coleta das informações necessárias para o envio dos questionários aos responsáveis: realizada através de busca na base de dados do CBBU e navegação nos *sites* das próprias bibliotecas e/ou instituições. As informações coletadas para o envio dos questionários encontram-se no Anexo 3;

4. envio dos questionários por e-mail;
5. recebimento dos questionários;
6. tabulação dos resultados mediante a elaboração de vários quadros; e
7. análise dos resultados.

Os resultados da etapa da análise empírica são mostrados no capítulo 6.

O trabalho inicia-se com esta introdução, na qual o tema é problematizado e justificado, e a pesquisa delimitada em termos de objetivos e metodologia.

No capítulo 2, são apresentados os fundamentos teóricos e técnicos do sistema de recuperação da informação: as atividades de indexação, classificação e catalogação e suas inter-relações são estudadas.

No capítulo 3, abordamos o conceito e evolução dos sistemas de recuperação da informação manuais e automatizados (*offline* e *online*), respectivas técnicas/métodos, além dos critérios utilizados para a avaliação de seu desempenho. Também são enfocados os instrumentos de recuperação da informação mais conhecidos: os esquemas de classificação bibliográfica, o tesauro e as listas de cabeçalho de assuntos.

No capítulo 4, são estudadas as principais questões relacionadas à recuperação da informação na *Web*, as especificidades do ambiente virtual, inclusive o fenômeno do crescente volume de documentos eletrônicos e das suas implicações no processo de recuperação da informação. Também analisamos mais especificamente na seção 4.1, a analogia entre metadados e catalogação e a utilização de metodologias "tradicionais" e/ou "convencionais" de bibliotecas no ciberespaço.

No capítulo 5, são analisadas as diversas interpretações, aplicações e atributos dos metadados e os tipos de entidades para descrição. Também apresentamos o que se constitui um esquema de metadados, nos detendo especialmente na sintaxe dos metadados. Por último, analisamos a interoperabilidade como fator de fundamental importância para o tema, destacando o papel desempenhado pelas *crosswalks* e *registries*.

No capítulo 6, apresentamos os resultados da análise empírica, conforme já descrito nesta introdução.

As conclusões e recomendações são abordadas no capítulo final.

O Anexo 1 apresenta o mapeamento dos esquemas de metadados no exterior, o Anexo 2 traz o questionário para a coleta de dados e, finalmente, o Anexo 3 apresenta as informações sobre as instituições coletadas.

2 FUNDAMENTOS TEÓRICOS E TÉCNICOS DO SISTEMA DE RECUPERAÇÃO DA INFORMAÇÃO: INDEXAÇÃO, CLASSIFICAÇÃO E CATALOGAÇÃO

Antes de abordarmos o sistema de recuperação da informação propriamente dito, é necessário enfocar as técnicas que lhes dão sustentação, ainda que de forma sucinta, mas com o objetivo de traçar sua evolução.

Segundo Lancaster (1979), o sistema de recuperação da informação tem como componentes: subsistemas de entrada (seleção de documentos, indexação e vocabulário) e subsistemas de saída (busca, comparação e interação entre o usuário e o sistema).

Subsistemas de entrada:

- subsistema de seleção de documentos: a entrada do sistema consiste em documentos que são selecionados de acordo com a política institucional, estabelecida a partir do conhecimento detalhado das necessidades de informação dos usuários do sistema;
- subsistema de indexação: organização e controle dos documentos adquiridos; as atividades de organização e controle incluem classificação, catalogação, indexação de assunto e resumo.
- subsistema vocabulário: escolha de "termos de indexação" de acordo com o vocabulário utilizado pelo sistema; podemos considerar que nesta etapa também ocorre a atividade de organização dos documentos e suas representações.

Subsistemas de saída:

- subsistema de busca: preparação de uma estratégia de busca pelos membros da equipe ("busca delegada") ou pelo próprio usuário ("busca não-delegada"), a partir do pedido realizado pelo usuário.
- subsistema de comparação ou confrontação (*match subsystem*): comparação entre as representações dos documentos e as representações das perguntas; e
- o subsistema de interação entre o usuário e o sistema (interface usuário-sistema): recuperação pelo sistema dos documentos que combinam com a estratégia de busca, sendo entregues ao solicitante.
- A estratégia de busca é constituída de duas etapas: a análise conceitual e tradução. A etapa de análise conceitual consiste na análise da pergunta para determinar o que realmente o usuário procura. A etapa de tradução envolve a tradução da análise conceitual no vocabulário do

sistema. A análise conceitual do pedido, traduzido na linguagem de busca é a "estratégia de busca".

O subsistema de indexação dos documentos é considerado por vários autores, dentre eles Lancaster (1979), um subsistema que impacta diretamente na capacidade em recuperar documentos que estejam de acordo com as necessidades de informação dos usuários do sistema. Este subsistema nos interessa particularmente para atender a um dos objetivos traçados por esta pesquisa, e por este motivo, analisamos as atividades de controle e organização que o caracterizam, quais sejam, a indexação, a classificação e a catalogação.

No Brasil, este tema é tratado por Robredo e Cunha (1986), que consideram a classificação, a catalogação e a indexação, técnicas de análise da informação e de representação do conteúdo dos documentos. Estas técnicas foram desenvolvidas devido à necessidade de localizar e recuperar a informação em grandes conjuntos de documentos, independente do tipo de suporte.

Há uma certa confusão na terminologia utilizada por diversos autores no que diz respeito às questões relacionadas à indexação dos documentos. Lancaster (1993) afirma que esta confusão tem origem em diferenças terminológicas bastante inexpressivas.

O objetivo da seção seguinte é tentar identificar, com maior clareza, as definições para os termos Indexação, Classificação e Catalogação, entendendo que estas técnicas de representação dos documentos não são estanques e sim inter-relacionadas, como veremos na seção 2.4. Para fins de apresentação e entendimento destes tópicos, cada uma destas técnicas será analisada separadamente, com o objetivo de definí-las e apresentar seus princípios, ainda que de forma sucinta. Também entendemos ser esta análise de grande importância para atigirmos um dos objetivos específicos desta pesquisa, que é estudar as relações dos metadados com as atividades de indexação, classificação e catalogação tradicional/convencional.

Cabe ainda acrescentar que é pertinente tecer um histórico da catalogação na seção 2.3, para melhor entendimento desta técnica, de forma a situar o surgimento de padrões importantes, reconhecidos internacionalmente, e ainda utilizados no mundo da catalogação "tradicional", tais como o AACR (Código Anglo-Americano de Catalogação), as ISBDs (Descrição Bibliográfica Internacional Normalizada), além de pontuarmos o aparecimento do formato MARC de catalogação, que inaugurou uma nova era nos sistemas de bibliotecas.

2.1 Indexação

A indexação consiste em identificar o assunto de que trata o documento, segundo Lancaster (1993). "Os termos atribuídos por um indexador servem como pontos de acesso mediante os quais um item bibliográfico é localizado e recuperado, durante uma busca por assunto num índice publicado ou numa base de dados legível por computador". (LANCASTER, 1993, p. 5)

Para Lancaster (1993), o processo de indexação implica na preparação de uma representação do conteúdo dos documentos e é considerado um dos fatores que determinam se uma base de dados é ou não bem sucedida. Para este autor, uma base de dados bem sucedida é aquela que consegue responder às indagações de seus usuários, que localiza os documentos que são úteis para satisfazer às suas necessidades de informação e que evita a recuperação de itens inúteis.

A indexação de assuntos consiste em duas etapas principais: a análise conceitual e a tradução. A análise conceitual é a etapa onde se decide o que trata o documento, isto é, a identificação do seu assunto ou assuntos. A etapa seguinte de tradução consiste na conversão da análise conceitual de um documento num determinado conjunto de termos de indexação (LANCASTER, 1993).

Na etapa de análise conceitual, na qual se decide o assunto ou assuntos do documento, devem ser consideraras as necessidades dos usuários do serviço, como bem aponta Lancaster (1993, p. 8):

> [...] uma indexação de assuntos eficiente implica que se tome uma decisão não somente quanto ao que é tratado num documento, mas também porque ele se reveste de um provável interesse para um determinado grupo de usuários. Em outras palavras, não existe um conjunto 'correto' de termos de indexação para documento algum. A mesma publicação pode ser indexada de forma bastante diferente em diferentes centros de informação e deve ser indexada de modo diferente, se os grupos de usuários estiverem interessados nesse documento por diferentes razões.

Outro aspecto importante na etapa de análise conceitual é a definição da política de indexação pelos administradores do sistema de recuperação da informação. Segundo Lancaster (1993), esta política se relaciona, fundamentalmente, com a exaustividade da indexação. No Brasil, Robredo e Cunha consideram que (1986, p. 246), "a exaustividade da indexação se refere ao nível de reconhecimento (e/ou) inclusão dos diferentes conceitos ou noções de que trata o documento". A grosso modo, segundo Lancaster (1993), a exaustividade pode ser considerada como o número de termos atribuídos ao item em média. Neste sentido, o autor estabelece uma distinção entre indexação exaustiva e indexação seletiva: a primeira corresponde ao emprego de

um número suficiente de termos de forma a contemplar o conteúdo do documento de modo bastante completo e, a segunda, ao emprego de um número muito menor de termos, de forma a abranger apenas o conteúdo temático principal. A exaustividade cresce a medida que aumenta o número de palavras presentes na representação de um item. Quando a indexação exaustiva é utilizada, ocorre alta revocação e menor precisão[3] de buscas, isto é, é recuperado um número maior de itens que o usuário considera não sendo pertinente à sua necessidade de informação. Já a indexação seletiva leva à maior precisão dos resultados. A quantidade de termos atribuídos ao documento constitui realmente uma questão de custo-eficácia: "quanto mais exaustiva for a indexação, maior será o custo e não é muito razoável indexar com um nível de exaustividade que não seja justificado pelas necessidades do usuário do serviço" (LANCASTER, 1993, p. 25).

Quanto à prática da indexação na etapa de análise conceitual, ao examinar o documento para identificar o que deve ser incluído na indexação, o indexador raramente poderá fazer uma leitura completa e estudo detalhado do item, devendo focar sua análise em partes do documento que apresentem "maior probabilidade de dizer o máximo acerca do conteúdo no menor tempo: o título, o resumo, o resumo do autor [*summary*] e as conclusões" (LANCASTER, 1993, p. 20). A este respeito, Robredo e Cunha (1986) salientam que normalmente a intenção do autor encontra-se estabelecida nos primeiros parágrafos, enquanto que as seções finais denotam o quanto os objetivos propostos foram atingidos pelo autor.

Lancaster (1993) aponta, ainda, algumas armadilhas a que o indexador está sujeito na prática da indexação, mais especificamente na identificação do que deve ser incluído na indexação: a) o indexador não deve ser influenciado pelo tipo de vocabulário a ser utilizado na etapa de tradução, em outras palavras, não pode ignorar um tópico porque acha que o mesmo não esteja contemplado adequadamente no vocabulário a ser utilizado; b) o indexador deve indexar as idéias do autor e não as palavras empregadas por ele, isto porque o autor pode estar utilizando termos que não estejam contemplados de forma exata no vocabulário controlado ou que apesar de serem exatamente iguais, tenham diferentes usos.

Na prática da indexação, um outro princípio descrito por Lancaster (1993, p. 27) como de fundamental importância e que remonta a Cutter, é o da especificidade, no qual "um tópico deve ser indexado sob o termo mais específico que o abranja completamente". Se não houver um termo sozinho que represente o conteúdo, pode-se buscar a combinação de termos. Nos sistemas de recuperação manuais, que antecederam os sistemas computadorizados, se fazia necessário o desdobramento das entradas dos termos específicos em termos mais genéricos, pois somente

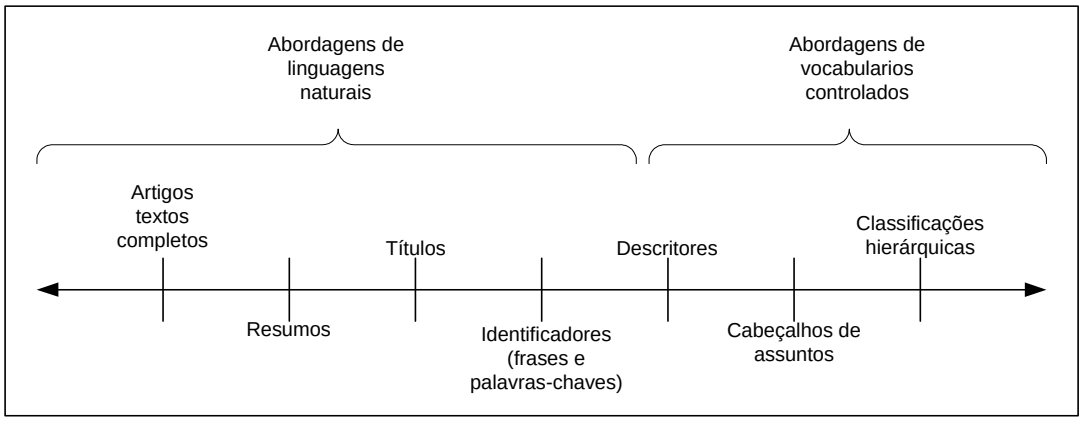
[3] Os termos precisão e revocação são conceituados na seção 3.2.

desta forma era possível realizar buscas mais genéricas, o que não é necessário em sistemas computadorizados e bem planejados, que utilizem um vocabulário controlado (LANCASTER, 1993).

A segunda etapa do processo de indexação, como já vimos anteriormente, é o processo de tradução, que envolve a representação da análise conceitual mediante a atribuição de um termo extraído de vocabulário. Este termo constitui um rótulo que identifica uma determinada classe de itens e pode ser uma palavra extraída de um tesauro, de uma lista de cabeçalhos de assuntos, do próprio documento, ou até mesmo extraída como um número de um esquema de classificação (LANCASTER, 1993). Neste sentido, Lancaster (1993) faz uma distinção entre indexação por extração (indexação derivada) e indexação por atribuição. Na primeira, palavras ou expressões são extraídas para representar o conteúdo temático dos documentos e pode ser também denominada indexação por palavra ou indexação livre (ROBREDO E CUNHA, 1986). Na indexação por atribuição, os termos que representam o assunto ou assuntos dos documentos são selecionados a partir de uma fonte que não é o documento: "mais freqüentemente, a indexação por atribuição envolve o esforço de representar a substância da análise conceitual mediante o emprego de termos extraídos de alguma forma de vocabulário controlado" (LANCASTER, 1993, p. 14). A indexação por atribuição também é denominada indexação por conceito, que "pressupõe a análise do conteúdo temático do documento (análise conceitual), a decisão sobre os conceitos presentes no texto e a tradução do observado em linguagem apropriada, com a qual rotulam-se os documentos e os seus registros bibliográficos" (PIEDADE, 1983, p. 10).

Portanto, na etapa de tradução, a indexação por extração utiliza a linguagem natural, enquanto a indexação por atribuição utiliza uma linguagem artificial que é controlada e codificada, ou também denominada de vocabulário controlado. A Figura 1 mostra as principais linguagens de descrição, denominadas também por alguns autores de linguagens documentárias.

Figura 1 - Linguagens de descrição da informação



Fonte: HARTER, Stephen P. Online Information Retrieval: concepts, principles and techniques. London: Academic Press, 1986, p. 42.

Do lado esquerdo são apresentadas as abordagens da linguagem natural para representação da informação, incluindo o texto completo, o artigo, o resumo e o título. A 4º classe denominada "identificadores" se refere a palavras-chaves extraídas do texto original pelos indexadores, utilizadas normalmente para complementar a indexação com palavras não representadas pelos termos constantes do tesauro (HARTER, 1986).

Ao lado direito temos os descritores (listados e apresentados no tesauro), os cabeçalhos de assuntos e as classificações hierárquicas. Estes tipos de vocabulários controlados serão analisados mais detalhadamente, ainda que de forma sucinta, na seção 3.3.

Há varios problemas que resultam da utilização da linguagem natural num sistema de recuperação da informação que podem ser resolvidos com a utilização de um vocabulário controlado, pois ao contrário da linguagem natural, este inclui, em geral, a forma de estrutura semântica que se destina especialmente a controlar sinônimos (estabelecendo uma única forma padronizada, com remissiva de todas as outras), a diferenciar homógrafos e reunir ou ligar termos cujos significados apresentem uma relação mais estreita entre si (relações hierárquicas e não-hierárquicas) (LANCASTER, 1993). A linguagem natural lida com palavras, e não conceitos, e um sistema de recuperação que a utilize "não permite e quase certamente não permitirá a busca efetiva de conceitos ou idéias diretamente" (HARTER, 1986, p. 31).

Desta forma, na utilização de um sistema baseado em linguagem natural, o usuário deverá antecipar todas as palavras e frases possíveis que poderão ser utilizadas para expressar o conceito de seu interesse. Harter (1986) observa que nas chamadas "ciências duras"[4] (*hard sciences*) o problema da ambigüidade semântica (como os homógrafos) tende a ser menor do que nas "ciências brandas" (*soft sciences*), pois nestas existe uma ambigüidade semântica inerente à própria disciplina.

O Quadro 1 mostra as características gerais da linguagem natural e dos vocabulários controlados. É importante frisar que os atributos listados como características dos vocabulários controlados não descrevem todos os vocabulários controlados de igual maneira. Estes atributos são generalizações e, enquanto tal, podem não ser aplicáveis para exemplos particulares (HARTER, 1986).

Quadro 1 - Características gerais da linguagem natural e dos vocabulários controlados

Linguagem Natural	**Vocabulários controlados**
Altamente expressiva	Não muito expressivo
Muito difícil executar buscas genéricas	Relativamente fácil executar buscas genéricas
Permite uma variedade de pontos de acesso	Permite apenas alguns pontos de acesso
Problema com sinônimos	Controle de sinônimos
Problema com homógrafos	Controle de homógrafos
Problema com *false drops*	Pré-coordenado para *false drops*
Altamente flexível	Altamente inflexível
Altamente representativa da realidade	Não muito representativo da realidade
Representa (quaisquer) muitos pontos de vista	Representa um único ponto de vista
Requer nenhum treinamento para ser utilizada	Requer treinamento
Fácil de representar novos conceitos	Difícil ou impossível representar novos conceitos
Fácil de representar conceitos complexos	Difícil ou impossível representar conceitos complexos
Ambígüa, *fuzzy* e branda	Sem ambiguidade, precisa e "dura"
Sem padronização	Padronizada
Liberdade de expressão	Altamente restritiva à liberdade de expressão
Não muito compacta	Altamente compacta
Indexação não necessária	Problemas de inconsistência na indexação
Usuário precisa pensar seus próprios Termos de busca	Termos adicionais sugeridos pela estrutura de referências cruzadas
Alto grau de exaustividade	Baixo grau de exaustividade

Fonte: HARTER, Stephen P. Online Information Retrieval: concepts, principles and techniques. London: Academic Press, 1986, p. 54.

[4] Como exemplo de "ciências duras" temos a física, a matemática, a química e as ciências naturais clássicas, em contraposição às "ciências brandas" que são as ciências sociais e humanas: "As "ciências duras" são denominadas paradigmáticas, isto é, seu corpo de crenças fundamentais, valores, suposições, atitudes e metodologias relacionadas a estrutura e identidade da disciplina, são compartilhadas pela comunidade de estudiosos e pesquisadores que trabalham naquela disciplina, enquanto as chamadas "ciências brandas" se encontram num estágio não-paradigmático, onde o consenso não existe e pode nunca vir a existir" (KUHN*apud* HARTER, 1986, p. 34).

Segundo Harter (1986), como pode ser visto no Quadro 1, tanto a linguagem natural como os vocabulários controlados apresentam vantagens e desvantagens para a indexação/recuperação da informação: uma é rígida, inflexível, mas precisa, a outra é altamente expressiva, flexível, mas potencialmente ambígüa. Isso leva o autor a concluir que há ocasiões em que a utilização da linguagem natural para indexação/recuperação será mais bem sucedida do que a utilização do vocabulário controlado e vice-versa. Segundo Harter (1986), "a maioria dos pesquisadores acredita que a "melhor" vertente, se ela existe, depende do problema da informação e seu contexto, assim como dos sistemas de busca e das bases de dados utilizadas" (p. 57).

Como vimos, o processo de indexação implica na preparação de uma representação do conteúdo dos documentos e, segundo Harter (1986), um registro indexado de um documento é uma representação do documento ou seu substituto, segundo um ponto de vista particular. No Brasil, em sua dissertação de mestrado, cujo título já é revelador, "Sistemas de redução da informação: uma (IR)Recuperação Metodologicamente Configurada", Pereira (1994, p. 69) levanta críticas ao sistema de recuperação da informação quando diz que "o processo de indexação consiste na geração de um modelo que passa, a partir de sua criação, a substituir o documento dentro do sistema", sendo considerado como um modelo que "desvia, esconde e mutila o universo de documentos que se propõe a representar".

2.2 Classificação

Há várias definições de classificação tanto de organizações como de pesquisadores da área. Podemos citar a definição dada pelo Comitê Técnico de Pesquisa em Classificação da *International Federation for Information and Documentation* (FID) (1973): "qualquer método de reconhecimento de relações genéricas ou outras, entre itens de informação, não importa o grau de hierarquia usada, nem se aqueles métodos são aplicados em conexão com sistemas tradicionais ou computadorizados" (CAMPOS, 2001, p.19). A *Enciclozyne*, uma enciclopédia digital na *Web*, define classificação como "o arranjo sistemático em grupos ou categorias de acordo com critérios estabelecidos".

Dentre os estudiosos da área, podemos citar a definição dada pela brasileira Piedade (1983, p. 16) em que "classificar é dividir em grupos e classes[5], segundo as diferenças e

[5] Classe é um conjunto de coisas ou idéias que possuem um ou vários atributos, predicatos ou qualidades em comum" (PIEDADE, p. 19). Podemos também utilizar a definição do Dicionário Online Dictionary for Information Science (ODLIS) em que classe "é um grupo de objetos ou conceitos baseados em uma ou mais características, atributos, propriedades, qualidades, etc., que possuem em comum, para o propósito de classificá-los de acordo com um sistema estabelecido, representado nos sistemas de classificação das bibliotecas por uma notação simbólica.

semelhanças. É dispor os conceitos segundo suas semelhanças e diferenças, em certo número de grupos metodicamente distribuídos". Para Lancaster (1993, p. 7), a classificação é uma atividade intelectual

> [...] que consiste em decidir do que trata um documento e de atribuir-lhe um rótulo que represente esta decisão, quer este rótulo seja extraído de um esquema de classificação, de um tesauro ou de uma lista de cabeçalhos de assuntos, [...]. No campo do armazenamento e recuperação da informação, a classificação de documentos refere-se à formação de classes de itens com base em seu conteúdo temático. Tesauros, cabeçalhos de assuntos e esquemas de classificação bibliográfica são essencialmente listas dos rótulos com os quais se identificam e, porventura, se organizam estas classes.

Os esquemas de classificação e as teorias que lhe são subjacentes serão objeto de análise na seção 3.3, considerando o papel desempenhado pelos esquemas de classificação enquanto ferramentas utilizadas para o fim último da recuperação da informação.

2.3 Catalogação

A terceira técnica é a catalogação, assim definida por Lancaster (1993) como o processo no qual o documento é identificado por elementos bibliográficos, tais como autores, títulos, fontes de publicação, etc, e outros dados julgados necessários. Segundo Gomes (1997, p. 1), "catalogação significa, em geral, descrição detalhada de objetos/peças de uma coleção". Ainda segundo a autora, no campo da Biblioteconomia, os objetos/peças são documentos e sua descrição pode se dar em dois planos diferentes: o da descrição física do documento (catalogação descritiva) e a descrição do assunto (catalogação de assunto). Como resultado desta atividade, temos o catálogo.

A catalogação é entendida por Robredo e Cunha como um processo de descrição bibliográfica, "onde todo documento é identificado por um número de registro, número de acesso ou número de amarração, além de outros elementos essenciais que o identificam, como o autor ou autores da obra considerada (livro, artigo de periódico, comunicação apresentada num congresso, etc), seu título e, conforme o caso, a imprenta ou a fonte, além de outros dados julgados necessários" (1986, p. 103).

Quanto aos seus objetivos e funções, segundo Cutter (*apud* BARBOSA, 1978, p. 23), o catálogo deve ser o instrumento que permita: a) encontrar um livro do qual se conheça o autor, o título ou o assunto, b) mostrar o que existe numa coleção de um determinado autor, ou sobre uma determinada obra".

Apesar da prática da catalogação pelos bibliotecários ser bem antiga, remontando às primeiras bibliotecas, a catalogação moderna tem seu início com a compilação de regras de catalogação para o Museu Britânico, por Anthony Panizzi, nos meados do século XIX. O que se

seguiu foi uma sucessão de códigos de catalogação, criados primeiramente por indivíduos influentes, como Charles Jewett e Charles Cutter e, *mais adiante, por organizações como a American Library Association (ALA) e a Library of Congress* (LC). Segundo Barbosa (1978), a história da normalização das regras catalográficas pode ser dividida, de maneira bem ampla, em três períodos distintos: a) de Panizzi[6] à Conferência de Paris[7], de 1841 a 1961, b) da Conferência de Paris à Reunião Internacional de Especialistas em Catalogação (RIEC), de 1961 a 1969 (período pré-mecanizado); e c) da RIEC ao Controle Bibliográfico Universal (CBU), de 1969 em diante (período mecanizado).

Segundo Barbosa (1978), no primeiro período temos a predominância de dois códigos de catalogação, a saber: o código da ALA (mais amplamente difundido na América) e as Instruções Prussianas (mais amplamente divulgadas na Europa). O código da ALA sofreu a influência e colaboração de Charles Ami Cutter, que consagrou a existência da escola americana de catalogação, ao publicar em 1876 a *Rules for a dictionary catalog*. Esta obra traz "369 regras que incluem normas não só para entradas por autor e por título, mas também para a parte descritiva, cabeçalhos de assuntos e ainda alfabetação e arquivamento de fichas" (BARBOSA, 1978, p. 29). Os princípios de descrição bibliográfica de Cutter influenciaram fortemente todos os códigos de catalogação que se seguiram. Um dos princípios mais conhecidos de Cutter é o da conveniência do usuário, que estabelece que o catálogo deve atender às necessidades de seus usuários, antes mesmo das necessidades do catalogador. Como decorrência deste, surgiu um outro princípio, o da facilidade do uso (BARBOSA, 1978).

O segundo período na história da normalização das regras catalográficas foi iniciado com a realização da Conferência de Paris, resultado de um movimento de reformulação das normas de catalogação até então utilizadas para atender às novas demandas surgidas no pós-guerra: com o avanço tecnológico deste período, houve o aparecimento de outros tipos de documentos em variadas formas de apresentação e conteúdo, causando um impacto considerável nos serviços de processamento técnico das coleções bibliográficas (BARBOSA, 1978).

Assim como a ALA em 1908, 1941 e 1949, as *Anglo-American Cataloging Rules* (AACR), publicadas pela primeira vez em 1967, sob os auspícios da LC, se baseou também nas regras estabelecidas por Cutter (CAPLAN, 2003).

Outro acontecimento importante neste período, apontado por Barbosa (1978), foi o aparecimento de um novo ator em cena, o computador, instrumento poderoso que começa a ser

6 Antony Panizzi foi autor do primeiro código de catalogação propriamente dito, publicado em 1839, cujas regras foram aprovadas em 1841, para utilização nos catálogos do Museu Britânico (BARBOSA, 1978).
7 Conferência Internacional de Princípios de Catalogação em Paris.

utilizado para muitos serviços realizados em bibliotecas, entre os quais a elaboração de catálogos. Para produzir o catálogo automatizado, o computador precisa de uma forma de interpretar a informação encontrada num registro catalográfico. Para atender esta necessidade, foi lançado em 1965 pela LC um projeto experimental denominado Projeto MARC I – *Machine Readable Cataloging* (catalogação legível por computador), linguagem padrão para a troca de informações bibliográficas, embrião dos programas de cooperação bibliotecária surgidos na década de 1970, como veremos mais adiante no período mecanizado da história da normalização (BARBOSA, 1978).

Na verdade, a LC já havia iniciado seus estudos sobre formatos bibliográficos legíveis por computador desde fins da década de 1950, com o objetivo de automatizar os processos de tratamento, armazenamento e recuperação de informações das grandes bibliotecas americanas. A partir do sucesso do projeto MARC I, houve um enorme esforço realizado internacionalmente para se chegar à padronização dos formatos para descrição bibliográfica, de forma a atingir um mínimo de entendimento entre sistemas para intercâmbio de seus registros bibliográficos em suporte magnético. Ao final do Projeto MARC I, em 1968, deu-se início ao desenvolvimento do MARC II, de concepção mais ampla, sendo adotado como padrão básico nacional americano para automação de processos técnicos em bibliotecas, utilizado pelas grandes redes prestadoras de serviços de informação nos Estados Unidos: a *Online Computer Library Center*[8] (OCLC), a *Western Library Network* (WLN) e o *Research Libraries Information Network* (RLIN) (BARBOSA, 1978).

Segundo a LC, no tutorial *Understanding Marc Bibliographic,* há razões importantes para a utilização de apenas um padrão: evitar a duplicação de trabalho, possibilitar melhor compartilhamento de recursos bibliográficos entre as bibliotecas. Outras razões não tão aparentes, mas igualmente importantes podem ser identificadas: atualmente, há muitos sistemas comerciais para o gerenciamento de bibliotecas de todos os tamanhos, desenhados para trabalharem com o formato MARC, sendo mantidos e melhorados pelos seus produtores para que as bibliotecas possam se beneficiar dos recentes desenvolvimentos tecnológicos. Além disso, o padrão MARC também permite que as bibliotecas possam substituir um sistema por outro, com a garantia de que seus dados ainda sejam compatíveis. O MARC passou por evolução para acompanhar as mudanças, sempre fiel ao esforço da integração de formatos e, atualmente, o formato utilizado é o MARC21, que será também abordado na segunda parte da dissertação, na seção 5.4.1.1.

[8] Denominada *Ohio College Library Center* até o ano de 1977 (PALMER, 1987).

O terceiro período da história da normalização iniciou-se com a RIEC, ocorrida em 1968, com o objetivo principal de conseguir em âmbito internacional, uma padronização da catalogação descritiva considerada imprescindível ao bom desempenho da catalogação compartilhada (*shared cataloging*), necessária para a disseminação da informação (BARBOSA, 1978). Como vimos, a partir do padrão MARC se confirmou a liderança da LC no campo da catalogação cooperativa, que tinha como finalidade acelerar a aquisição e a catalogação de livros e a aplicação do computador em bibliotecas. Segundo Carvalho (1999, p. .22), "boa parte da literatura publicada a partir da década de 60 trata a cooperação bibliotecária a partir do surgimento das redes de bibliotecas, redes de informação e da automação das bibliotecas".

O formato criado pela LC para registrar seus dados bibliográficos converteu-se pouco depois numa norma do *American National Standards Institute* (ANSI), que veio a ser recomendada pela *International Organization for Standardization* (ISO), como norma internacional, a ISO 2709: *Documentation Format for Bibliographic Interchange on Magnetic Tape*, publicada em 1973, que seria revista posteriormente em 1981, para transformar-se em referencial para todos os formatos de intercâmbio de informações atualmente no mundo inteiro (CAPLAN, 2003).

Robredo e Cunha (1986) salientam que a ISO 2709 é um formato de comunicação e intercâmbio e não um formato para processamento interno pelos diversos sistemas. Para o intercâmbio de informações, os sistemas podem utilizar os formatos internos que desejarem, com a condição de respeitar algum tipo de padrão que permita a conversão do formato interno em formato de comunicação e intercâmbio, e vice-versa.

Neste mesmo período surgiu o Controle Bibliográfico Universal (CBU), criado pela Unesco, um programa a longo prazo para controle e permuta de informações bibliográficas em âmbito internacional. No início da década de 1970, a *Internacional Federation for Library Associations* (IFLA) desenvolveu uma série de regras denominadas ISBD[9]: *Internacional Standard Bibliographic Description* (Descrição Bibliográfica Internacional Normalizada), com o objetivo de encorajar a padronização da prática de catalogação, internacionalmente. Várias especificações ISBD foram elaboradas, das quais podemos citar: ISBD (G): *General International Standard Bibliographic Description* (Descrição Bibliográfica Internacional Normalizada Geral); ISBD (M): *International Standard Bibliographic Description for Monographic Publications* (Descrição Bibliográfica Internacional Normalizada para Monografias). A versão de 1988 do código de catalogação Anglo-Americano

[9] Uma lista atualizada da família das ISBDs está disponível em: http://www.ifla.org/VI/3/nd1/isbdlist.htm. Acesso em: 28.07.04.

(AACR2R) foi resultado de uma revisão substancial, baseada amplamente nas ISBDs. (CAPLAN, 2003).

Embora alguns códigos sejam reconhecidamente menos bem sucedidos do que outros, as regras de catalogação bibliográfica sempre tentaram manter-ser fiéis a princípios fundamentais, incluindo o princípio da conveniência do usuário, sempre tentando facilitar os objetivos do catálogo (CAPLAN, 2003). Segundo a autora, os objetivos delineados por Cutter, há mais de 100 anos atrás, estão refletidos hoje nas primeiras três tarefas do usuário definidas na *IFLA Functional Requirements for Bibliographic Records* (Requisitos Funcionais para Registros Bibliográficos da IFLA):

- Encontrar entidades[10] que correspondam aos critérios de busca estabelecidos pelos usuários (localizar um única entidade ou um conjunto de entidades num arquivo ou base de dados como resultado de uma busca utilizando um atributo ou relação da entidade);
- Identificar uma entidade (confirmar que a entidade descrita corresponda à entidade procurada ou distinguir entre duas ou mais entidades com características similares);
- Selecionar uma entidade que seja apropriada às necessidades do usuário (escolher uma entidade que satisfaça as necessidades do usuário no que se refere ao conteúdo, formato físico, etc. ou rejeitar uma entidade por ser imprópria às necessidades do usuário); e
- Adquirir ou obter acesso à entidade descrita (adquirir uma entidade através da compra, empréstimo, etc., ou acessar uma entidade eletronicamente através de uma conexão *online* a um computador remoto).

2.4 Inter-relações entre Indexação, Classificação e Catalogação

As técnicas abordadas na seção anterior não são estanques e sim inter-relacionadas, como bem demonstram as análises dos autores apresentadas a seguir.

Alguns especialistas fazem uma distinção entre catalogação de assuntos e indexação de assuntos, a primeira sendo as atribuições de cabeçalhos de assuntos para representar o conteúdo total de itens bibliográficos completos (livros, relatórios, periódicos, etc) e, a segunda, correspondendo às atribuições de cabeçalhos de assuntos[11] para partes de itens bibliográficos completos (artigos de periódicos, capítulos de livros). Para Lancaster (1993), esta distinção é "artificial, enganosa e incongruente" (p. 15). Neste sentido, Gomes (1997) afirma que "de um modo geral, podemos considerar neste contexto como termos equivalentes catalogação de

[10] Os tipos de entidades da IFLA serão analisadas na seção 5.3.
[11] Os cabeçalhos de assuntos serão analisados na seção 3.3.3.

assuntos e indexação de assuntos, porquanto são processos muito semelhantes, com diferenças adjetivas" (p. 1).

A mesma confusão se dá na distinção entre catalogação de assuntos e classificação. A primeira é considerada o ato de atribuir ao documento o cabeçalho de assunto e, a segunda, a atribuição do número de classificação. Sobre esta distinção, Gomes afirma que "na verdade, organizar os assuntos dos documentos reunindo-os segundo aspectos comuns é o mesmo que classificar" (1997, p. 1). Robredo e Cunha (1986) salientam que a diferença entre estes dois processos atenua-se ao se pensar que ambos têm o mesmo objetivo de identificar a informação com vistas à sua localização e recuperação. "Na própria etimologia dos dois termos (classificação – do latim, ação de fazer classes – e catalogação – do grego, ação de subdividir o conhecimento), encontramos a raiz de uma mesma preocupação, a partir de duas abordagens diferentes, de ordenar as informações ou os conhecimentos, ou seus suportes, juntando-os por grupos ou classes que guardam certa afinidade, para localizá-los dentro de um conjunto mais amplo" (ROBREDO e CUNHA, 1986, p. 202).

No Brasil, outra autora ratifica esta idéia: "o processamento técnico da informação, constituído essencialmente pela catalogação, classificação e indexação é indissociado, tanto que alguns autores consideram a classificação (esquemas de classificação, universais e especializados) parte das linguagens de indexação, ao lado de listas de termos (cabeçalhos de assunto), listas de descritores e tesauros" (PINHEIRO, 2002, p.7).

Até então, os autores citados são especialistas das áreas de Biblioteconomia e Ciência da Informação. Mas especialistas de outras áreas da informação, como Schellenberg (1980, p. 335), autor clássico da Arquivologia, também aborda o tema e apresenta as diferenças entre índices e catálogos, produtos das atividades de indexação e catalogação, respectivamente:

> Há porém, entre eles, diferença de grande importância e relativa, principalmente, ao modo como neles se identificam os documentos. No catálogo, tal se faz mediante o fornecimento de dados sobre o responsável pela produção, o tipo, o lugar, a data desta e sua quantidade. Nos índices, os documentos se identificam tão-somente pelo símbolo ou pelo nome do produtor. Neles, outrossim, indica-se apenas o conteúdo dos materiais e, de ordinário, nenhuma informação biográfica ou bibliográfica é proporcionada. A distinção entre índices e catálogos deriva dos fins a que se destinam. Conceberam-se os primeiros exclusivamente para permitir o acesso ao assunto – mera indicação de onde se pode encontrar, nos documentos, informação sobre os tópicos. Não visam, como os catálogos, a descrição dos papéis, mas simplesmente caracterizá-los em relação aos temas. Os índices representam, pois, meios de localização, ao passo que os catálogos são instrumentos descritivos, embora, como é óbvio, lhes seja dado servir para situar a informação pertinente.

3 SISTEMA DE RECUPERAÇÃO DA INFORMAÇÃO

Neste capítulo abordaremos o conceito e evolução dos sistemas de recuperação da informação manuais e automatizados (*offlline e online*), respectivas técnicas/métodos, desde a década de 1940 até os dias atuais, além das medidas utilizadas para avaliação de desempenho dos sistemas de recuperação da informação. Serão enfocados também os instrumentos de recuperação da informação, quais sejam, os esquemas de classificação bibliográfica, o tesauro e a lista de cabeçalhos de assuntos, mencionados anteriormente, enquanto o papel de destaque desempenhado pelas atividades de classificação, catalogação e indexação já foi objeto de análise do capítulo anterior. No entanto, antes de analisarmos a evolução do sistema de recuperação da informação, é pertinente inserí-lo na Ciência da Informação.

Ao citar Wersig e Nevelling que atribuem à Ciência da Informação a responsabilidade social de transmitir conhecimento para os que necessitam, Saracevic (1992, p. 9) enfatiza seu caráter social e conceitua a área como

> um campo dedicado à investigação científica e prática profissional que trata dos problemas de efetiva comunicação de conhecimentos e de registros do conhecimento entre seres humanos, no contexto de usos e necessidades sociais, institucionais e/ou individuais de informação.

O desenvolvimento da Ciência da Informação como campo científico e profissional se devem para Saracevic (1992), em grande parte, aos resultados alcançados no desenvolvimento de produtos, sistemas, redes e serviços na recuperação da informação. A evolução da Ciência da Informação está intrinsicamente ligada às questões relacionadas aos sistemas de recuperação da informação: muitos dos esforços e recursos da área foram e ainda são gastos para solucionar os problemas associados aos sistemas de recuperação da informação. A recuperação da informação não é a única atividade na Ciência da Informação, mas a maior fonte de relações interdisciplinares. (SARACEVIC, 1992).

Na introdução deste trabalho, Otlet e Bush já foram citados como precursores da Ciência da Informação. As origens da recuperação da informação e da própria Ciência da Informação remontam a Paul Otlet, documentalista belga, que em sua obra *Traité de Documentation*, escrito em 1934, nos brindou com idéias revolucionárias para o seu tempo, como o *Mundaneum*, um centro internacional para armazenamento e disseminação do conhecimento (RIEUSSET-LEMARIÉ, 1998). Mas Paul Otlet não foi o único a ter idéias inovadoras, outro precursor foi Vannevar Bush, idealizador de um máquina de recuperação da informação imaginária e que, em artigo intitulado *As We May Think*, escrito em 1945, propõe o MEMEX, baseado na noção de associação, o

mesmo padrão que o cérebro humano utiliza para assimilar informação. Ao criar este aparato, Bush objetivava solucionar os problemas advindos da explosão informacional, fenômeno característico do pós-guerra, resultante dos esforços de pesquisa desenvolvidos durante a segunda guerra mundial. Neste contexto, "a visão de Bush não é a única mas partilhada por uma série de cientistas que começaram a se dedicar à criação de métodos de organização e acesso a conjuntos de informação, tendo em vista não mais seu armazenamento mas sua reutilização" (NOVELLINO, 2000, p. 43).

Bush entendia que o conhecimento só poderia ser utilizado se fosse selecionado e recuperado. Neste sentido, outro teórico fundamental para a Ciência da Informação foi Borko, que nos mostra o quão importante é a pesquisa na área, ao investigar as propriedades e comportamento da informação, a utilização e a transmissão da informação, bem como o processamento da informação para armazenagem e recuperação ótimas. Borko (1968, p. 3) define a Ciência da Informação como uma área

> [...] interessada num conjunto de conhecimentos relacionados com a origem, coleção, organização, armazenagem, recuperação, interpretação, transmissão, transformação e utilização da informação. Inclui a investigação das representações da informação nos sistemas naturais e artificiais, a utilização de códigos para transmissão eficiente da mensagem, o estudo de instrumentos e técnicas de processamento da informação, tais como computadores e seus sistemas de programação. É uma ciência interdisciplinar derivada e relacionada com a matemática, a lógica, a lingüística, a psicologia, a tecnologia do computador, a pesquisa operacional, as artes gráficas, as comunicações, a biblioteconomia, a administração e assuntos similares. Tem componentes de uma ciência pura, que investiga o assunto sem relação com sua aplicação, e componentes de uma ciência aplicada, que cria serviços e produtos.

O termo Recuperação da Informação foi criado por Mooers em 1951, que o definiu como uma operação que "abarca os aspectos intelectuais de descrição da informação e sua especificação para a busca, e também quaisquer sistemas, técnicas ou máquinas que sejam empregadas para realizar esta operação" (MOOERS *apud* SARACEVIC, 1992, p. 7).

Entre os autores mais relevantes, na área de recuperação da informação, destaca-se Lancaster (1979), cuja abordagem privilegia o sistema de recuperação da informação freqüentemente citado no capítulo anterior. Para este autor, a principal finalidade deste sistema é assegurar que a necessidade de informação de um membro da comunidade de usuários seja atendida na hora em que ele necessite. Na sua definição, a recuperação da informação é um processo de busca de um conjunto de documentos, termo por ele adotado em sentido amplo, de forma a identificar os documentos relativos a um assunto em particular. Qualquer sistema que é empregado para facilitar esta atividade de busca de literatura (*literature search*) pode ser legitimamente chamado de sistema de recuperação da informação. Mas o próprio autor faz a ressalva de que este termo, apesar de ser amplamente utilizado, não é satisfatório para descrever o

tipo de atividade para a qual é normalmente aplicado, pois "um sistema de informação não recupera informação, já que informação é alguma coisa intangível. Somos 'informados' sobre um assunto se o nosso estado de conhecimento sobre este assunto é de algum modo modificado, pois informação é algo que muda o estado de conhecimento de alguém sobre um determinado assunto" (LANCASTER, 1979, p. 12). É oportuno aqui, relembrar as idéias de Pereira (1994), que também apresenta um outro olhar sobre o sistema de recuperação da informação, sobre a (IR)Recuperação, abordado anteriormente.

3.1 Sistemas de recuperação da informação e sua evolução

A evolução dos sistemas de recuperação da informação dependem muito dos avanços obtidos nas técnicas e métodos empregados com este objetivo, ilustrados por Saracevic (1992, p. 3) através de "exemplos históricos" que demonstram a evolução da área:

> [...] de cartões perfurados para sistemas *online* e CD-ROM, de sistemas sem capacidades interativas para aqueles que oferecem interações múltiplas, munidos de interfaces inteligentes, transformando a recuperação da informação em um processo altamente interativo, de bases de documentos para bases de conhecimento, de textos escritos para multimídia, de recuperação da citação para recuperação do texto completo, e até mesmo para sistemas especializados e de pergunta/resposta (*question answering*) e assim por diante.

Palmer (1987) é outro autor que se dedica ao estudo dos sistemas de recuperação da informação, mais especificamente, os sistemas *online*. Em seu livro, *Online Reference and Information Retrieval*, ele traça como principal objetivo da obra, oferecer ao profissional da informação um panorama sobre os sistemas de recuperação da informação *online*, tais como o ORBIT, o DIALOG e WILSONLINE. Logo na introdução, PALMER delineia a evolução dos sistemas de recuperação da informação a partir da década de 50, evolução esta marcada pelo impulso da automação, decorrente da incorporação e desenvolvimento dos computadores para o processamento de grandes volumes de dados. Portanto, da mesma forma que Lancaster, ele considera os computadores como "agentes de mudança, indispensáveis no processo de armazenamento e recuperação do conhecimento" (1987, p. 1)

Os "agentes da mudança" de Palmer também passaram por inúmeras transformações e desenvolvimentos com o decorrer dos anos, como atestam Robredo e Cunha (1986, p. 25) ao demonstrarem as sucessivas gerações de computadores, diferenciadas umas das outras pelos diferentes componentes físicos utilizados na memória central do computador:

> [...] nos computadores antigos, se utilizavam válvulas eletrônicas, que eram ligadas e desligadas para representar as codificações de *bits*. Nos anos 60 e início da década de 70, os computadores usavam nas suas memórias circuitos com núcleo de metal, passíveis de serem magnetizados (os conhecidos núcleos de ferrite). As memórias dos novos computadores atuais geralmente armazenam as informações em circuitos eletrônicos,

> em vez de circuitos magnéticos. As memórias são compostas de microscópicos circuitos integrados de silício (geralmente conhecidos como *chips*) ou outros materiais semicondutores. Assim, cada caractere é representado pela presença ou ausência de corrente elétrica numa determinada combinação de circuitos.

Apresentaremos a seguir, os vários desenvolvimentos dos sistemas de recuperação da informação. Escolhemos traçar esta evolução por década, assim como outros autores também o fizeram, para que possamos apresentar com clareza as principais características e eventos de cada um dos períodos.

3.1.1 Década de 40

Antes da década de 40, segundo Lancaster (1993), o sistema de recuperação mais rudimentar era um catálogo de fichas utilizado em bibliotecas. Nestes sistemas manuais de recuperação que antecedem os sistemas computadorizados, o processo de indexação tinha como produto o índice impresso ou o catálogo em fichas, denominados sistemas pré-coordenados. Lancaster delineia as principais características destes sistemas: "1. É difícil representar a multidimensionalidade das relações entre os termos, 2. Os termos somente podem ser listados numa determinada seqüência (A, B, C, D, E), o que implica que o primeiro termo é mais importante que os outros, 3. Não é fácil (senão completamente impossível) combinar termos no momento em que se faz uma busca" (1993, p. 42).

Diferentemente dos índices pré-coordenados, os índices pós-coordenados, que surgiram na década de 40, apresentam maior flexibilidade. Para Lancaster, "a recuperação da informação eficiente demanda sistemas que permitam a "combinação" livre de classes e termos que as representam – e os índices pré-coordenados baseados em entradas lineares não permitem a combinação de termos". (1979, p. 20). Ainda segundo este autor (1993), as principais características destes sistemas são: "1. Os termos podem ser combinados entre si de qualquer forma no momento em que se faz a busca, 2. Preserva-se a multidimensionalidade das relações entre os termos, 3. Todo termo atribuído a um documento tem peso igual: nenhum é mais importante que o outro" (LANCASTER, 1979, p. 33).

Lancaster faz um paralelo entre os sistemas pós-coordenados e os sistemas *online*, ao se referir a estes "como um descendente direto destes sistemas manuais". (1993, p. 32).

3.1.2 Década de 50

Na década de 50, apesar da existência de computadores, estes ainda apresentavam uma série de limitações: alto custo, disponibilidade limitada, velocidade de processamento lenta e pequena memória interna para manipular dados. Além disso, para que os computadores fossem

utilizados, eram exigidos altos níveis de habilidade técnica por parte dos seus usuários. Nesta época, as principais mídias de armazenamento utilizadas eram os cartões perfurados ou fitas magnéticas para processamento seqüencial de dados (PALMER, 1987).

Nessa década, o sucesso do *Sputnik I*, primeiro satélite artificial lançado ao espaço pela União Soviética, fez com que os Estados Unidos, receosos de que estivessem ficando para trás tecnologicamente, percebessem a necessidade de melhorar a eficiência da transferência de informação científica, o que acarretou investimento na pesquisa em recuperação da informação (BELLCORE, 1995).

O índice KWIC (*keyword in context*) [palavra-chave no contexto] surgiu nessa década. Segundo Lancaster (1993), é um método simples de produção de índices impressos por computador, que trabalha com textos e principalmente com as palavras que ocorrem nos títulos dos documentos. No índice KWIC, utilizado por pesquisadores como H. P. Luhn, é destacada cada palavra-chave que aparece no título no centro da página, sendo envolvida pelas palavras restantes do título. Lancaster aponta que "o programa de computador que gera o índice identifica as palavras-chave mediante um processo 'reverso': ele reconhece as palavras que não são palavras-chave (constantes de uma lista de palavras proibidas) e impede que sejam adotadas como pontos de entrada. As palavras desta lista têm função sintática (artigos, preposições, conjunções, etc.), mas, em si, não indicam conteúdo temático" (1993, p. 48).

O índice KWIC é um instrumento barato utilizado para se obter um certo nível de acesso temático ao conteúdo de uma coleção (LANCASTER, 1993). Ele se "tornou popular rapidamente por ser um meio não-trabalhoso, rápido e de baixo custo, de prover acesso por assunto à informação técnica" (PALMER, 1987, p. 02). Segundo Bellcore (1995), a lógica do KWIC: "qualquer ocorrência de qualquer palavra", ainda sobrevive atualmente como tipo de processamento em muitos sistemas de recuperação comerciais.

3.1.3 Década de 60

O final da década de 50 e início da década de 60 é considerado um período de grande experimentação na área da recuperação da informação. Segundo Bellcore (1995), datam desta época a construção do primeiro sistema de informação em larga escala, a elaboração das definições de revocação (*recall)* e precisão (*precision)*, o desenvolvimento da tecnologia para avaliação dos sistemas de recuperação da informação e a separação do campo da Recuperação da Informação do ramo principal da Ciência da Computação. Palmer (1987, p. 3) corrobora as idéias de Bellcore (1995) sobre a década de 60 quando se refere a este período como de pesquisa básica

intensa nos Estados Unidos, quando os efeitos residuais do *Sputnik* incitaram a utilização de fundos federais para as bibliotecas e para a pesquisa na área da informação:

> A teoria da informação e o crescimento do conhecimento estavam entre os tópicos estudados. Fatores humanos no desenho dos sistemas, assim como o comportamento de usuários eram considerados. A aquisição e a representação da informação receberam especial atenção. Houve revisões sobre indexação, resumos, classificação, codificação, estruturas de arquivo e estratégias de busca. Diretrizes para medidas de avaliação de sistemas e serviços foram criadas.

Dentre os muitos trabalhos deste período, o autor destaca o livro escrito por Tefko Saracevic, *Introduction to Information Science*, que contém 65 artigos representando o amplo espectro de pesquisa na área e surgimento do *Annual Review of Information Science and Technology* (ARIST), publicação que compila artigos de revisão organizados em tópicos.

Este período foi marcado pelo desenvolvimento das bibliotecas e sistemas de informação, decorrente dos custos decrescentes e maior disponibilidade de *hardware*, avanços tecnológicos e de rede de comunicação de dados: "Soluções centralizadas e em larga escala para a aquisição, catalogação, controles de periódicos, circulação e empréstimos entre-bibliotecas na universidade e em grandes bibliotecas públicas se tornaram o foco do desenho dos sistemas". (PALMER, 1987, p. 2)

Nessa época, mais precisamente no ano de 1966, foram realizados estudos de viabilidade que concluíram pela necessidade de reformatação dos registros em padrão MARC (*Machine-readable catalog*) [catalogação legível por computador] para uso em bibliotecas locais, o primeiro formato de intercâmbio de dados criado para a catalogação informatizada.

O surgimento de padrões permitiu que bibliotecas de todos os tipos e tamanhos pudessem compartilhar e utilizar os dados catalogados através de serviços bibliográficos, tais como a *Online Computer Library Center* (OCLC), fundada em 1967 por iniciativa das universidades no Estado de *Ohio* para desenvolver um sistema computadorizado, no qual as bibliotecas das instituições acadêmicas deste estado americano pudessem compartilhar recursos e reduzir custos. O seu primeiro presidente, Frederick G. Kilgour, vislumbrou a transformação da OCLC de âmbito regional para uma rede cooperativa internacional. Atualmente, a OCLC serve a mais de 45.000 bibliotecas de todos os tipos nos Estados Unidos e em 84 países e territórios por todo o mundo. Iniciativas como a da OCLC, segundo Palmer (1987), tornaram-se operacionais com o desenvolvimento dos computadores que deram suporte a recursos de processamento multi-usuário (*time-sharing)*: conectados a um computador *online* por linha de telefone, que podiam ter acesso ao computador de terminais remotos, tudo a um custo baixo, como veremos na década de 1970.

Foi nos Estados Unidos, na década de 60, que os sistemas de recuperação da informação baseados em computador surgiram e seu processamento era *offline*, segundo Lancaster (1979). Este autor menciona, entre as instituições pioneiras do processamento bibliográfico por computador em larga escala, a Biblioteca Nacional de Medicina, nos Estados Unidos, através da base de dados MEDLARS, lançada em 1963, que indexa artigos biomédicos. O MEDLARS é um dos maiores sistemas de informação e foi um dos primeiros a se tornar disponível em larga escala. Segundo Palmer (1987), F. Wilfrid Lancaster, com suas pesquisas, contribuiu para que a MEDLARS tivesse um padrão de qualidade excepcional, e cita dois livros que foram um marco desta época, frutos do trabalho de Lancaster na Biblioteca Nacional de Medicina: o *Information Retrieval Systems* e o *Vocabulary Control for Information Retrieval.*

O sistema de recuperação da informação em computador trouxe uma série de vantagens, das quais podemos citar (LANCASTER, 1979, p. 67):

> 1. possibilidade de realizar diversas buscas ao mesmo tempo;
>
> 2. habilidade em prover muitos pontos de acesso a um documento, de maneira extremamente econômica;
>
> 3. habilidade em lidar com buscas complexas envolvendo um número grande de termos e suas complexas relações;
>
> 4. habilidade em gerar uma saída (*output*) na forma de bibliografia impressa;
>
> 5. habilidade em coletar, de forma sistemática, dados de gerenciamento sobre o funcionamento do sistema;
>
> 6. habilidade em produzir muitas saídas (*outputs*) e serviços de uma única operação de entrada (*input*);
>
> 7. possibilidade de duplicar a base de dados de forma simples e barata, com o objetivo de ser utilizada na provisão de serviços de informação por um número diferentes de centros.

Quanto às suas características, esses sistemas eram muito parecidos entre si: o processamento era *offline*, conforme mencionado anteriormente, utilizando a fita magnética como mídia de armazenamento e a busca era seriada. A maioria dos sistemas era baseada na indexação humana e no uso de estratégias de busca preparadas por humanos, atividades apoiadas por um vocabulário controlado. Mas estes sistemas apresentavam uma série de desvantagens: eram sistemas de uma só tentativa (*one-chance*), onde o usuário tinha que pensar antecipadamente em todas as possibilidades de busca. Além disso, não era possível obter resposta imediata a uma consulta e o usuário precisava delegar a responsabilidade pela preparação de uma estratégia de

busca a um especialista da informação. A maioria destes sistemas oferecia tanto a Busca Retrospectiva quanto a Disseminação Seletiva da Informação[12] (LANCASTER, 1979).

3.1.4 Década de 70

Nesta década, é importante mencionar o desenvolvimento dos computadores: em janeiro de 1975, surge o primeiro PC (*Personal Computer*), o Altair 8800, veiculado em artigo publicado no periódico *Popular Electronics*, baseado no microprocessador INTEL 8080; e em 1979 surge o Apple II.

Durante a década de 60, o número de bases de dados havia crescido de menos de 100 para mais de 600 (PALMER, 1987). Alguns dos fatores responsáveis por este crescimento: o papel preponderante dos produtores das bases de dados ao estabelecerem redes ou outras atividades de cooperação, num nível nacional ou internacional; o surgimento do *Scientific Information Dissemination Center* (SIDC) (para a Disseminação Seletiva da Informação) e do Centro de Serviço *Online* (mais Busca Retrospectiva do que Disseminação Seletiva da Informação), que funcionavam como serviços intermediários entre o produtor de bases de dados e o usuário final; e o reconhecimento gradual de que qualquer cientista ou outro profissional pudesse acessar qualquer base de dados que necessitasse, na hora desejada. Mas o fator mais importante foi o aparecimento das habilidades de busca *online*, que tornaram as bases de dados amplamente acessíveis, representando uma verdadeira revolução na provisão de sistemas de informação (LANCASTER, 1979).

Os sistemas bibliográficos *online* existiam, pelo menos de forma experimental ou como protótipo por quase 15 anos, mas somente no final da década de 60, foi disponibilizado o primeiro sistema de Recuperação *Online* de larga escala: o *Remote Console Information Retrieval Service* (RECON), construído por *Lockheed Missiles & Space Company* para a *National Aeronautics and Space Administration* (NASA), cujo desenvolvimento começou em 1965, tornando-se operacional apenas em 1969 (LANCASTER, 1979).

Os Sistemas de Informação *Online*, diferentemente dos sistemas *offline*, são heurísticos e interativos, permitem o *browsing*[13] (navegação), além de serem capazes de fornecer uma resposta rápida. Nestes sistemas, o usuário pode fazer a busca diretamente, sem a intermediação de um especialista de informação (LANCASTER, 1979).

---

[12] A Busca Retrospectiva se dá em todos os documentos da base de dados, enquanto que a Disseminação Seletiva da Informação (DSI) ocorre apenas nos documentos recém-acrescentados ao sistema, uma vez utilizados para propósitos de DSI, serão da mesma forma acrescentados à base de dados permanente, sendo mais tarde também utilizados para a busca retrospectiva.

[13] A utilização do termo *browser* é anterior a *Internet* e foi traduzido pela expressão "folheando a esmo".

Uma outra característica do Sistema *Online* é o processamento multi-usuário (*time-sharing)*, em que o tempo de processamento do computador se divide entre duas ou mais atividades independentes, permitindo que diferentes usuários tenham acesso ao sistema ao mesmo tempo, criando a ilusão de que o usuário de cada terminal é o único a desfrutar das facilidades oferecidas pelo computador. Outra característica importante é a operação em tempo real (*real-time*), em que o computador recebe os dados, processa-os, retornando rapidamente os resultados, em tempo suficiente para que sejam utilizados numa atividade em pleno andamento. Na maioria das vezes, um sistema *online* bem desenhado pode responder a uma pergunta ou comando tão rapidamente, que a resposta é caracterizada como quase imediata (LANCASTER, 1979). Palmer (1987) aponta o processamento multi-usuário (*time-sharing)* como um dos principais fatores para o desenvolvimento da área de Recuperação da Informação na década de 70, pois possibilitou uma recuperação muito mais prática, pois agora as respostas dadas pelo sistema eram quase que imediatas.

A Biblioteca Nacional de Medicina (*National Library of Medicine*) evoluiu para oferecer aos seus usuários um sistema *online*, através do software ORBIT, lançando o MEDLINE (MEDLARS *Online*) em 1971. Nele, os usuários podiam efetuar buscas em um nível de profundidade e complexidade que estava além da capacidade de índices impressos ou de outras ferramentas manuais (LANCASTER, 1979).

Em 1972, a *Lockheed Missiles & Space Company* lançou o Serviço de Recuperação da Informação DIALOG. E em 1973, Carlos Cuadra, utilizando o software ORBIT (o mesmo utilizado pela *National Library of Medicine* para o MEDLINE), supervisionou a implementação do SDC ORBIT *Search Service* (PALMER, 1987).

Como uma alternativa ao alto preço cobrado pelos serviços *online* comerciais, foi lançado em 1977, por Janet Egeland e Ronald Quake, ambos oriundos do *Biomedical Communication Network* (BCN) da Universidade do Estado de Nova York, o Sistema de Informação BRS, que fornecia acesso mais barato ao MEDLINE e a um número pequeno de bases de dados (PALMER, 1987).

As décadas de 70 e 80 são consideradas por Bellcore (1995) como um período de pesquisa em bases de dados e em automação de escritórios. Houve pesquisa na área de recuperação da informação, mas não como na década de 60, parte disso se deveu a uma reorientação da política do Governo dos Estados Unidos. Mas, ainda assim, houve algum progresso na área e o mais importante avanço foi o aparecimento da recuperação da informação probabilística, liderada por Keith van Rijsbergen. Esta pesquisa trouxe novas técnicas como a

medição de freqüência de palavras em documentos relevantes e não-relevantes, a utilização de medidas de freqüência de termos para ajustar o peso dado a diferentes palavras (BELLCORE, 1995).

3.1.5 Década de 80

Palmer (1987, p. 5) se refere aos desenvolvimentos tecnológicos contínuos na área de circuitos integrados que na década de 80 tornaram os PCs e seus componentes menos caros e mais poderosos:

> [...] os dois milhões de caracteres (*bytes*) da memória principal e os 30 milhões de caracteres (30 MB) de armazenamento em disco que estavam disponíveis apenas para grandes centros de computador a menos de uma década passada, são agora lugar comum entre as instalações de computadores pessoais. Palavras de tamanho maior, memórias maiores e *softwares* mais integrados (para processamento simultâneo da base de dados, comunicação e processamento de palavras) estão tomando lugar de aplicações complexas dentro do alcance das menores bibliotecas com os orçamentos mais restritos. Os fabricantes líderes de computadores – Apple com o Macintosh II e a IBM com o Sistema / 2 máquinas – concorrem entre si na introdução de novos *softwares* e *hardwares*.

Um novo ator surge neste cenário, uma mídia de armazenamento muito mais poderosa que o disco flexível: o CD-ROM. Esta nova mídia começa a ser utilizada para distribuir informação. Com capacidade de armazenamento bem maior, custo baixo e facilidade de uso, o CD-ROM foi utilizado para armazenar bases de dados, como o DIALOG, que distribuiu o *Dialog OnDisc*. (PALMER, 1987)

A rede de computadores continuou a se desenvolver nesta década, mas o CD-ROM se enquadrava tão bem para a publicação da informação tradicional, que se desenvolveria como uma ameaça aos sistemas *online* que estavam crescendo rapidamente por duas décadas (BELLCORE, 1995).

Durante a década de 80, o crescimento regular do *word processing* e a diminuição dos preços do espaço em disco significou que mais e mais informação estava sendo disponibilizada na forma legível por máquina e que era mantida desta forma (BELLCORE, 1995). O uso dos sistemas de recuperação *online* se expandiu em dois caminhos principais: disponibilização de textos completos ao invés de apenas resumos e indexação e a expansão de sua utilização por não-especialistas, pois as bibliotecas substituíram ou complementaram seus catálogos em fichas pelo acesso público de seus catálogos. Um exemplo foi a *Library of Congress* com o *REMARC project*.

Palmer (1987) aponta o surgimento de um novo termo, o hipertexto, que "descreve sistemas de bases de dados experimentais que permitem que o usuário passe de um documento para outro através de *links*. Na leitura de um artigo de enciclopédia, por exemplo, um usuário

pode pressionar um botão para ir diretamente para a seção de um documento suporte (*supporting document*), ao invés de seguir por um tedioso caminho de referências cruzadas e notas de pé de página". (PALMER, 1987, p. 6). O termo hipertexto, na verdade, foi criado por Theodor (Ted) Nelson na década de 60, quando ele construiu a visão de Xanadu, influenciado em grande parte pelas idéias de Vannevar Bush. Como já vimos, Bush introduziu a noção de associação de conceitos ou palavras na organização da informação, baseado no padrão que o cérebro humano utiliza para assimilar informação em conhecimento.

Técnicas e diferentes aspectos dos sistemas de recuperação da informação relativos à *Internet*/*Web* correspondem à década de 1990, quando o uso da rede realmente se consolidou e se ampliou, no mundo inteiro, inclusive no Brasil. Assim, estas questões serão abordadas especificamente no capítulo 4, que enfoca a recuperação da informação na *Web*.

3.2 Critérios de avaliação dos sistemas de recuperação da informação

Para a apresentação dos critérios de avaliação dos sistemas de recuperação da informação, nos baseamos apenas em Lancaster, por entender que este autor foi um dos primeiros a estabelecer tais critérios de forma sistematizada, tendo outros autores sempre nele se fundamentado para apresentar análises igualmente valiosas.

O desempenho de um sistema de recuperação da informação pode ser medido pela satisfação do usuário em ter suas necessidades de informação atendidas e para sua avaliação Lancaster (1979) estabeleceu os seguintes critérios:

- revocação;
- precisão;
- cobertura;
- esforço do usuário; e
- tempo de resposta.

A taxa de revocação é definida como a habilidade do sistema em recuperar documentos relevantes. A revocação é a relação entre o número de documentos relevantes recuperados e o total de documentos relevantes existentes na base de dados (recuperados e não recuperados) (LANCASTER, 1979).

A taxa de precisão se refere à habilidade do sistema em evitar documentos irrelevantes e corresponde a relação entre o número de documentos relevantes recuperados e o número total de documentos recuperados (relevantes e não relevantes). Para estes dois critérios são

normalmente inversamente proporcionais. Quanto maior a revocação, menor a precisão e vice-versa. Mas, dependendo da necessidade do usuário, o melhor desempenho pode ser obtido com uma alta taxa de revocação ou com uma alta taxa de precisão (LANCASTER, 1979).

O conceito de relevância na literatura, segundo Lancaster (1979), está atrelado ao julgamento de um indivíduo ou grupo de indivíduos, portanto, é impreciso e variável. Um documento é relevante quando ele é *julgado relevante* por aquele que fez a busca, atendendo à sua necessidade de informação, estando suficientemente próximo do assunto solicitado.

Segundo o mesmo autor, o critério de cobertura é uma extensão da revocacão, expressado em termos da quantidade de literatura que uma base de dados tem sobre um determinado assunto. Este critério é particularmente importante para quem precisa fazer uma busca exaustiva sobre um determinado assunto.

O critério relativo ao esforço do usuário é medido pelo tempo gasto por ele para conduzir sua busca, o quanto de esforço é feito para que ele utilize o sistema e aprenda a usá-lo (treinamento do usuário), na interpretação da forma em que os resultados de busca são apresentados e na obtenção dos documentos descritos (LANCASTER, 1979).

O quinto e último critério de avaliação de performance de um sistema de recuperação da informação é o tempo de resposta, que é diferente para uma busca intermediada e uma busca não-intermediada. No primeiro caso, é o tempo gasto entre a submissão do pedido pelo usuário e o acesso aos resultados da busca. Já no segundo, é o tempo envolvido na condução da busca, e, neste caso, também é uma medida de esforço do usuário (LANCASTER, 1979).

3.3 Instrumentos de recuperação da informação

Com o objetivo de entender melhor as particularidades do esquemas de classificação bibliográficas, do tesauro e dos cabeçalhos de assuntos, cada um destes instrumentos de recuperação da informação será analisado separadamente, a seguir, ainda que de forma sucinta, uma vez que já foram tratados em capítulo anterior.

3.3.1 Esquemas de classificação bibliográfica

A classificação bibliográfica é analisada por Campos enquanto "esquema que permite a organização e a recuperação do conhecimento registrado" (2001, p. 28). Muitos autores fazem uma distinção entre a classificação bibliográfica e a classificação filosófica. Em sua análise, Piedade (1983) apresentou os pontos de vista de vários teóricos da área a respeito de suas diferenças e similaridades. Ainda segundo a autora, a classificação bibliográfica é aquela que tem

por base os assuntos tratados nos documentos" (1983, p. 65). As classificações filosóficas, segundo Piedade (1983, p. 61) são

> [...] criadas por filósofos com a finalidade de definir e hierarquizar o conhecimento. Surgiram quando os sábios compreenderam que o universo é um sistema harmônico, cujas partes estão dispostas em relação ao todo, que há uma hierarquia das causas e dos princípios e, portanto, uma hieraquia e uma relação entre as ciências que as estudam e resolveram esquematizar estas hierarquias, criando as classificações filosóficas.

Campos (2001) baseia-se também nos mesmos teóricos apontados por Piedade para explanar a respeito da dupla função da classificação bibliográfica: a de permitir a organização dos documentos nas estantes e a de representar o conhecimento registrado numa dada área de assunto.

As classificações bibliográficas, em virtude das características próprias aos documentos, além das divisões do conhecimento, exigem, segundo Piedade (1983):

- uma classe que reúna as obras sobre todos os assuntos, subdividida pela forma do documento;
- subdivisões de forma, aplicáveis aos vários assuntos; e
- uma notação, isto é, um conjunto de símbolos para representarem os assuntos e permitir a ordenação lógica dos documentos.

A classificação bibliográfica envolve o desenvolvimento e utilização de um esquema de classificação. Campos (2001) entende como fundamental, para a compreensão dos esquemas de classificação, a análise das teorias de classificação bibliográfica que são subjacentes a estes esquemas.

A teoria da classificação bibliográfica passou por dois estágios de evolução: o primeiro estágio é o da teoria descritiva e o segundo, o da teoria dinâmica (KUMAR *apud* CAMPOS, 2001). Até a década de 30, os esquemas de classificação bibliográficos existentes não eram flexíveis a ponto de absorverem novos assuntos em suas tabelas, tornando-se rapidamente obsoletos. Estes primeiros esquemas denominados descritivos eram organizados

> [...] a partir dos assuntos representativos da literatura da área, naquele momento histórico, isto é, os elementos constitutivos dos esquemas são os assuntos representados a partir da freqüência de ocorrência na literatura. Só permitem, por isso mesmo, representar o conhecimento já estabelecido. Daí a dificuldade em classificar assuntos novos, muitos dos quais ainda sem um nome fixado. Pode-se afirmar que, naqueles esquemas, não ocorre a ligação entre o conhecimento e as classificações, mas entre os assuntos dos documentos e as classificações (CAMPOS, 2001, p. 32).

Neste sentido, ao desenvolver a Teoria da Classificação Facetada na década de 1930, Ranganathan estava consciente da necessidade de elaborar esquemas de classificação que

pudessem acompanhar as mudanças e evolução do conhecimento. Ranganathan é um dos primeiros teóricos que, ao explicar a natureza da classificação bibliográfica, percebeu a necessidade de elaborar esquemas de classificação que pudessem acompanhar as mudanças e a evolução do conhecimento, através de sua Teoria Dinâmica do Conhecimento. Para Campos (2001, p. 33), a diferença entre a Teoria Descritiva e a Teoria Dinâmica repousa no fato de que

> [...] o assunto não está pronto no esquema, ele é construído no momento da análise do documento. Assim, se o uso da Teoria Descritiva permite representar o conhecimento registrado de um dado momento histórico, a Teoria Dinâmica, por sua vez, vai interagir com esta realidade, já que possui princípios que norteiam a elaboração de esquemas flexíveis.

O próprio Ranganathan classificou os esquemas descritivos em: Esquema de Classificação Enumerativo (a *Library of Congress Classification* e a *Rider´s International Classification*): "consiste numa única tabela, que relaciona todos os assuntos passados, presentes e futuros" (PIEDADE, 1983, p. 67), Esquema de Classificação Quase Enumerativo (*Decimal Classification* de Mevil Dewey e a *Subject Classification* de J. D. Brown): "consta de longas tabelas enumerativas para a maioria dos assuntos, acompanhadas de algumas tabelas de subdivisões comuns" (PIEDADE, 1983, p. 68) e Esquema de Classificação Quase Facetado (*a Universal Decimal Classification* e a *Bibliographic Classification* de J. Bliss): "compõe-se de tabelas enumerativas de assuntos, completadas por tabelas de subdivisões especiais" (PIEDADE, 1983, p. 68).

O primeiro esquema de classificação facetado baseado na teoria dinâmica do conhecimento é a *Colon Classification* de Ranganathan[14]. Segundo Campos (2001), as edições posteriores da *Colon Classification* apresentam aperfeiçoamentos que levam Ranganathan a classificar as primeiras edições do seu esquema de classificação em Rigidamente Facetados e as posteriores em Livremente Facetados (ou Analítico-Sintéticos). Os sistemas rigidamente facetados "são constituídos de tabelas contendo assuntos básicos, tabelas de subdivisões comuns, tabelas auxiliares especiais e determinações rígidas sobre a seqüência em que devem ser combinados os vários conceitos (fórmula-de-facetas)" (PIEDADE, 1983, p. 68). Os sistemas livremente facetados ou analítico-sintéticos apresentam as mesmas partes que o tipo anterior, mas não determinam a ordem para a combinação dos vários conceitos, passando esta combinação a ser guiada por princípios, possibilitando ao classificador criar novas subdivisões, segundo normas estabelecidas" (PIEDADE, 1983, p. 68).

[14] Classificação de Dois Pontos.

3.3.2 Tesauro

O tesauro é um vocabulário controlado que surgiu na década de 60, como um instrumento de indexação/recuperação, controlando aspectos semânticos e lingüísticos, de forma a contribuir para um disciplinamento do vocabulário usado na indexação de serviços bibliográficos. A primeira e mais simples forma de vocabulário controlado é o uso de descritores, que se encontram listados e descritos num tesauro. Normalmente, é originado de uma coleção dinâmica e crescente de documentos, em que os elementos do vocabulário possuem relações lógicas uns com os outros (HARTER, 1986). As relações básicas entre os elementos do vocabulário em um tesauro são de equivalência, hierárquica e de associação ou afinidade. Para representar estas relações utilizam-se as expressões:

- BT *broader term* (termo mais amplo, mais genérico)
- NT *narrower term* (termo mais estreito, mais específico)
- USE *use* (use)
- UF *used for* (usado no lugar ou usado para)
- RT *related term* (termo relacionado)
- SN *scope note* (nota de escopo ou nota de abrangência)

As expressões BT a NT são utilizadas para sugerir relações hierárquicas que põem em evidência as relações de subordinação genérico-específico dos termos. A utilização das expressões USE e USED FOR apontam para a escolha de um termo preferido para ser utilizado como descritor, disciplinando o problema dos sinônimos da linguagem natural. A expressão RT sugere que dois conceitos estão de alguma forma relacionados um com o outro, sendo que esta relação não pode ser hierárquica (uso do BT e NT), nem tampouco de sinonímia (USE, UF). Segundo Robredo e Cunha (1986), a relação de associação pode ser de diversos tipos: antonímia (oposição), coordenação, descendência, concorrência, causa-efeito e instrumental. Por último, a SN é utilizada para clarear o significado pretendido, se há mais de um uso potencial de uma palavra numa base de dados (distinção entre homógrafos).

3.3.3 Lista de cabeçalhos de assuntos

A Lista de Cabeçalhos de Assuntos é "uma lista alfabética completa de um vocabulário controlado criado por catalogadores e utilizado na catalogação desde 1898 pela Biblioteca do Congresso para designar cabeçalhos de assunto para facilitar o acesso ao conteúdo da informação

dos trabalhos publicados", de acordo com o *Dicionário Online Dictionary for Information Science* (ODLIS).

Segundo Campos (2001), "o tesauro veio a se contrapor às listas de cabeçalhos de assuntos" (p. 90). Estas listas adotam uma terminologia mais geral do que a que encontramos no tesauro. Poucos são os termos sugeridos como relacionados, através da utilização do *see also* (ver também) e *see also from* (ver também de). A indicação *see also* não distingue as relações hierárquicas e termos relacionados, diferentemente do tesauro que utiliza BT, NT e RT para fazer estas distinções. Como exemplos de Listas de Cabeçalhos de Assuntos temos aquelas mais utilizadas pela maioria das bibliotecas nos Estados Unidos: *Sears List of Subjsct Headings* e a *Library of Congress Subject Headings* (LCSH).

Mas há também uma importante diferença filosófica entre o tesauro e os cabeçalhos de assuntos. Os cabeçalhos de assuntos são baseados em coleções específicas de documentos. Ao contrário, o tesauro é derivado de coleções de livros, revistas, etc, existentes e crescentes, relativas à uma área. O vocabulário no tesauro é utilizado para resolver os problemas de sinonímia e ambigüidade semântica nestas coleções.

4 A RECUPERAÇÃO DA INFORMAÇÃO NA *WEB*

O catálogo composto de descrições estruturadas de objetos de informação pode ser encontrado sempre quando temos grandes coleções de objetos que precisam ser gerenciados. Podemos conceituar objeto de informação como: "um item ou grupo de itens digitais, seja qual for o tipo ou formato, que pode ser localizado ou manipulado como um objeto único por um computador" (GILLILAND-SWETLAND, 1998, p. 5).

A importância do catálogo cresce na mesma proporção do tamanho da coleção a ser descrita. E um dos grandes problemas apontados por vários autores é a não existência de um catálogo que possa gerenciar a que é considerada, sem sombra de dúvida, a maior coleção de objetos do mundo, a *World Wide Web* (GILL, 1998).

A *Web* atualmente apresenta um volume maciço de informações. Para melhor entendermos esta explosão informacional, nos baseamos no estudo realizado anualmente por Lyman e Varian (2003), intitulado *How Much Information*, que analisa as taxas de crescimento e o fluxo de informações em várias mídias, dentre elas, a *Internet*. Segundo o estudo, embora a *Internet* seja a mais nova das mídias[15], é a que cresce com maior rapidez. O estudo faz distinção entre a *Web* de superfície (*surface Web*), que representa a fração da *Web* de acesso público e gratuito e a *Web* profunda (*deep Web*), também denominada *Web* invisível (*hidden Web*), que se refere à fração da *Web* cujas páginas só existem como resultado de buscas nas bases de dados[16]. A *Web* de superfície perfaz o volume de 167 *terabytes*, enquanto a *Web* profunda está na faixa de 91,850 terabytes. Em 2000, o volume estimado de informações na *Web* era de 20 a 50 terabytes, e em 2003 o volume atingiu 67 terabytes, portanto, no período de 2000 a 2003, segundo esse estudo, o volume de informação na *Web* de superfície triplicou. Outro dado impressionante é que, no mundo inteiro, cerca de 600 milhões de pessoas têm acesso a *Internet*.

Ao analisar o problema da oferta excessiva de dados, do excesso e falta de informação, Froelich (1998, p. 2) se refere à *Web* como uma anti-coleção: "é uma miscelânea de itens, surgindo numa diversidade de formas, com pouca ou nenhuma autoridade ou controle, com pouca organização global, mecanismos de busca bem pobres". O autor entende esta anti-coleção como um paradoxo, pois apesar da proliferação de materiais, há uma extraordinária carência de informação na *Web*.

---

[15] Além da *Internet*, o estudo considera também as seguintes mídias: rádio, televisão e telefone.

[16] Ler também o artigo: BERGMAN, Michel K. The deep *Web*: surfacing hidden value. Journal of the Electronic Publishing. V. 7, n. 1, Aug. 2001. Disponível em: http://www.press.umich.edu/jep/07-01/bergman.html. Acesso em: 15/04/04.

Infelizmente, nem a *Web* e nem a *Internet* – a infra-estrutura de redes, servidores e canais de comunicação que lhe dão sustentação – foram originalmente desenhadas com a idéia da catalogação de seus conteúdos. O protocolo TCP/IP, que permite o funcionamento da infra-estrutura básica da *Internet* é uma camada de transporte, para a transferência rápida e segura de pacotes de dados de um ponto ao outro, enquanto que o *Hyper Text Transfer Protocol* (ou HTTP) lida apenas com a entrega de informação através de *links* na *World Wide Web*. Isso significa que os protocolos existentes na rede não oferecem nenhum suporte para a localização específica de recursos de informação (GILL, 1998). Souza e Alvarenga (2004, p. 3), em ensaio sobre a *Web* Semântica e suas contribuições para a Ciência da Informação, confirmam este estado de coisas:

> Embora tenha sido projetada para possibilitar o fácil acesso, intercâmbio e a recuperação de informações, a *Web* foi implementada de forma descentralizada e quase anárquica; cresceu de maneira exponencial e caótica e se apresenta hoje como um imenso depositório de documentos que deixa muito a desejar quando precisamos recuperar aquilo de que temos necessidade. Não há nenhuma estratégia abrangente e satisfatória para a indexação de documentos nela contidos, e a recuperação das informações, possível por meio dos "motores de busca" (*search engines*), é baseada primeiramente em palavras-chaves contidas nos textos dos documentos originais, o que é muito pouco eficaz.

Com o crescimento de páginas HTTP e com o objetivo de solucionar o problema da localização de recursos de informação, os serviços conhecidos atualmente como mecanismos de busca (*search engines*)[17] começaram a aparecer (SCHWARTZ, 1998). Estas ferramentas surgiram logo após o aparecimento dos primeiros *browsers* (navegadores), como o lançado pela *European Organization for Nuclear Research* (CERN), no início da década de 90, e as versões gráficas dos navegadores para *Windows* e *Macintosh,* em 1993. Dentre as primeiras ferramentas podemos citar a *WWW Virtual Library*, fundada por Tim Berners-Lee em 1992, pouco tempo depois do lançamento da própria *Web,* e da *Webcrawler, Yahoo!* e *Lycos*, lançadas em 1994 (GILL, 1998).

Os mecanismos de busca disponíveis atualmente para ajudar os usuários a encontrar recursos na *Web* são maiores e mais potentes que os seus predecessores e precisam ser para que possam acompanhar a explosão do crescimento, tanto de informação disponível, quanto de usuários acessando a *Web* (GILL, 1998). A maioria dos autores identifica duas classes principais de mecanismos de busca: os diretórios e os motores de busca.

Os diretórios são formados por listas hierárquicas de *sites*, subdivididos em categorias e subcategorias. Os *sites* passam por um processo de seleção, realizado por seres humanos, que estão sempre atualizando o diretório ao descobrirem novos recursos por meio de sugestões de usuários, por pesquisas na própria *Web*, ou até mesmo utilizando robôs para localizar novas

[17] O termo *search engines* é traduzido em português por mecanismos de busca ou ferramentas de busca.

URLs[18]. Os diretórios podem ser genéricos como a *World Wide Web Virtual Library* e o *Yahoo!* e o brasileiro Cadê, ou podem ser especializados em áreas de assunto particulares, tais como o *Art, Design, Architecture & Media Information Gateway* (ADAM) e o *Edinburgh Enginnering Virtual Library* (EEVL). Os diretórios fornecem acesso aos seus *links* mediante a busca ou navegação no conjunto hierárquico de cabeçalhos de assunto (GILL, 1998).

Segundo Cendón (2001), os diretórios foram a primeira solução para organizar e localizar recursos na *Web*, numa época em que seu conteúdo ainda era pequeno o suficiente para permitir que fosse coletado de forma não automática, tendo precedido os chamados motores de busca. Como vimos, a *World Wide Web Virtual Library* foi o primeiro mecanismo de busca do tipo diretório lançado na *Web*.

Diferentemente do diretório, o motor de busca não organiza suas páginas de forma hierárquica e utiliza o método de robôs. Na verdade, o motor de busca é formado por quatro elementos: o robô[19] que varre a *Web* na busca por documentos; um indexador, que extrai a informação das páginas HTML e as armazena numa base de dados; a interface, normalmente uma página *Web* que é utilizada pelos usuários da ferramenta para realizar a pesquisa na base de dados, e, por último, o motor de busca propriamente dito, que mediante a busca solicitada, localiza dentre os milhões de itens da base de dados, aqueles que devem constituir uma resposta. O motor de busca é um programa que também é responsável pela ordenação dos resultados, de maneira que os mais citados apareçam no topo da lista (CENDÓN, 2001).

Segundo Céndon (2001), ao contrário dos diretórios, os motores de busca não organizam hierarquicamente as páginas que colecionam. Preocupam-se menos com a seletividade que com a abrangência de suas bases de dados, procurando colecionar o maior número possível de recursos. Conseqüentemente, suas bases de dados são extremamente grandes, podendo alcançar centenas de milhões de itens. A busca é baseada em palavras-chaves (*keywords*), ou, às vezes, em linguagem natural.

Já os motores de busca surgiram quando o volume de informações na *Web* começou a crescer assustadoramente, tornando a coleta por meios manuais e a busca através da navegação muito difíceis. Os primeiros motores de busca baseados em palavras-chaves foram o *Archie-Like Indexing on the Web (AliWeb)* e o *Harvest*, que utilizavam tecnologias diferentes dos motores de busca atuais, enquanto que o *WebCrawler*, lançado em abril de 1994, foi o primeiro motor de

[18] *Uniform Resource Locator.*
[19] O robô também é chamado aranha (*spider*), rastejador (*crawler*), viajante (wanderer) e ainda verme (*worm*).

busca baseado em robô, tecnologia utilizada atualmente por todos os motores de busca (CENDÓN, 2001).

Vários autores consideram uma terceira classe de mecanismos de busca, além dos diretórios e motores, as chamadas metaferramentas. Estas ferramentas permitem a execução de uma mesma busca em mais de um mecanismo de busca, apresentando ao usuário todos os resultados numa única lista. Na verdade, as metaferramentas não possuem nenhuma base de dados, apoiando-se nas bases de dados dos mecanismos de busca. Cendón (2001) apresenta como exemplo deste tipo de mecanismo de busca, as seguintes metaferramentas: *Dogpile*, *Savvy Search* e *Mamma*.

Contudo, há sérios problemas de ambas visões dos diretórios e motores de busca. Se por um lado, os diretórios especializados oferecem mais precisão nos resultados das buscas, constituindo-se num ambiente na rede que armazena coleções de informação de maior qualidade, conseqüência da intervenção humana nos processos de indexação e classificação, por outro lado, esta mediação é um processo custoso que demanda muito trabalho e tempo e não consegue oferecer cobertura ampla de toda a *Web*, por conta do volume desmensurado de informações e pela própria natureza temporária dos recursos nela disponíveis. Outra questão importante no que se refere, ainda, à catalogação dos recursos de informação da *Web* por seres humanos, é decidir o nível de detalhamento da descrição, que vai depender largamente da sua finalidade e da maior ou menor importância do recurso a ser descrito para o serviço de informação (GILL, 1998).

Os motores de busca, por sua vez, também apresentam problemas relativos principalmente à capacidade da ferramenta em manter um índice de páginas de cobertura ampla e atualizada e à pouca probabilidade em encontrar o que se procura, mesmo que tenha sido indexado pelo motor. GILL (1998) nos apresenta alguns destes problemas:

- Os componentes dos motores de busca são totalmente automatizados, o que significa que os recursos da *Web* são selecionados por *software* e não por pessoas, sendo variáveis em qualidade.
- A busca em bases de dados muito extensas, indexadas automaticamente, sempre resultam em conjuntos de resultados extremamente numerosos, muito freqüentemente não aproveitados pelos usuários, a despeito das ferramentas de recuperação da informação serem cada vez mais sofisticadas, da aplicação de procedimentos de relevância e da utilização de algoritmos de inteligência artificial que levem em conta o contexto (*context-aware*).
- Os motores não conseguem indexar as páginas geradas como resultados de busca nas bases de dados, a parcela que corresponde a *Web* invisível, o que é no mínimo preocupante, já que

há grande quantidade de informações sendo geradas nesta fração da *Web,* como demonstrado anteriormente pelo estudo de Lyman e Varian (2003).

- A largura de banda[20] da *Web*, exigida pelos motores de busca para manter índices atualizados e abrangentes, pode alcançar níveis inaceitáveis devido ao aumento do volume de informação.

Ainda segundo GILL (1998), embora os diretórios e os motores de busca sofram de uma série de problemas, uma análise cuidadosa demonstra que a maioria das dificuldades é resultado de ambições insustentáveis a longo prazo: o fato é que a *Web* está se tornando muito grande para que uma só organização ou serviço possa ser capaz de catalogá-la, não importando se utilizam pessoas ou computadores para gerar seus índices.

Uma das soluções preconizadas para o problema da descoberta de recursos na *Web* é a proposta de algum tipo de catálogo distribuído. GILL aponta a *WWW Virtual Library* como um exemplo que, apesar dos esforços altruísticos de seus curadores voluntários, foi insuficiente para acompanhar o crescimento da *Web* (GILL, 1998).

Para a construção deste catálogo distribuído, pelo menos em nível técnico, a interoperabilidade já não é mais um problema, pois protocolos técnicos como o Z39.50 já estão disponíveis[21]. O que é necessário, agora, são os padrões mais abstratos para a estrutura e conteúdo da informação que permita a interoperabilidade em nível semântico (GILL, 1998). E esta é justamente a visão da *Web* Semântica, um projeto do *World Wide Web Consortium* (W3C) que pretende operar uma transformação na *Web* como a conhecemos hoje.

Segundo Souza e Alvarenga (2004), devemos entender a conotação "semântica" para a *Web* como atrelada a idéia de estabelecer associações dos documentos a seus significados através de metadados descritivos. É neste contexto que devemos situar as "ontologias", construídas em consenso pelas comunidades de usuários e desenvolvedores de aplicações, de forma a permitir o compartilhamento de significados comuns. Segundo Souza e Alvarenga (2004, p.4)

> [...] o projeto da *Web* Semântica, em sua essência, é a criação e implantação de padrões tecnológicos para permitir a construção desta nova *Web*, que não somente facilite as trocas de informações entre agentes pessoais, mas que principalmente estabeleça a língua franca para o compartilhamento mais significativo de dados entre dispositivos e sistemas de informação de uma maneira geral.

20 Capacidade de transportar informações.
21 O protocolo Z39.50 é abordado mais especificamente na seção 5.5

A língua franca a que se referem os autores é o Dublin Core. Por esta razão, ao estudarmos os esquemas de metadados, enfocaremos este padrão em contraponto ao formato de catalogação bibliográfica mais utilizado pelas bibliotecas no mundo todo, o MARC, mostrando comparativamente as características gerais de ambos. Apesar do Dublin Core ser a língua franca, é importante notar que não há consenso sobre o melhor esquema de metadados, apesar dos esforços realizados mundo afora neste sentido. Já existem centenas de esquemas de metadados e este número está crescendo rapidamente em função das diferentes comunidades e necessidades de seus membros. Assim, como atestam Milstead e Feldman (1999), qualquer grupo pode começar seu próprio esforço de definição de metadados para atender a seus interesses específicos. As autoras entendem esta profusão de padrões como o maior empecilho ao desenvolvimento ordenado de metadados, referindo-se a esta situação como uma "atmosfera caótica de padrões". Com isso, vamos apresentar a importância das *crosswalks* e dos *registries* para a interoperabilidade entre vários sistemas, baseados em diferentes esquemas de metadados.

Ainda sobre o projeto da *Web* Semântica, Souza e Alvarenga (2004) advogam que é fundamental a padronização de tecnologias, de linguagens e de metadados descritivos: os usuários da *Web* devem obedecer a regras comuns e compartihadas sobre como armazenar dados e descrever a informação armazenada para que esta informação possa ser "consumida" por outros usuários humanos ou não, de maneira automática e não-ambígua. E acrescentam que "o primeiro passo para este objetivo está sendo a criação de padrões para a descrição de dados e de uma linguagem que permita a construção e codificação de significados compartilhados". (SOUZA e ALVARENGA, 2004, p. 4)

É com este objetivo que estudaremos as várias linguagens de marcação ou sintaxes existentes, inclusive a recomendada pelo próprio W3C, a linguagem XML e que, segundo muitos, será a linguagem do futuro na *WWW*.

4.1 Catalogando sob um outro nome ...[22]

Antes de mais nada, achamos importante pontuar a discussão a respeito das duas vertentes de pensamento que advogam diferentes estratégias para organizar a *Internet*: uma que entende que é responsabilidade das instituições atuais a tarefa de catalogar e organizar materiais digitais e outra que acredita que novas ferramentas e técnicas farão desnecessárias a necessidade do uso de métodos "tradicionais" (WOODWARD, 1996).

[22] O título desta seção é inspirado no título do trabalho de Milstead e Feldman (1999), "Cataloging by Any Other Name ...", citado nas referências bibliográficas.

Neste tópico, procuramos mostrar também que os metadados podem ser entendidos como uma nova aplicação para as técnicas de representação do conteúdo dos documentos, tão conhecidas e utilizadas pelos bibliotecários por décadas. Esta analogia deve ser entendida dentro da concepção de que no ciberespaço técnicas e metodologias "tradicionais" ou "convencionais" de bibliotecas, tais como catalogação, classificação e indexação, estão sendo utilizadas na estruturação da informação e organização do conhecimento, transportas, de forma atualizada, adaptada e expandida. (PINHEIRO, 2002).

A despeito desta discussão, é fato que o *expertise* tradicional da Biblioteconomia está se traduzindo em uso efetivo na *Internet*, e, para demonstrar tal fato, fazemos também uma breve explanação sobre algumas iniciativas na *Web* em que estas "velhas" práticas são utilizadas. Antes de mais nada, precisamos definir o que consideramos como "tradicional" ou, ainda, "convencional". Neste sentido, Woodward (1996) entende que não se pode "congelar" um corpo de conhecimento e experiência no tempo e que a Biblioteconomia tem evoluído por um longo período. A autora conceitua como "tradicional", "aquelas técnicas desenvolvidas principalmente no final do século XIX e no século XX que são utilizadas quase que exclusivamente em bibliotecas e sistemas de indexação" (WOODWARD, 1996, p. 190).

Quando o termo metadados começou a ser utilizado na *Internet* e na *Web,* no contexto da descrição de objetos de informação na rede, os bibliotecários foram rápidos em perceber que metadados eram apenas um novo nome para uma prática já conhecida e utilizada por eles há bastante tempo, a catalogação. Na verdade, os profissionais de informação têm utilizado o termo ao se referirem ao ato de catalogar ou indexar informações que eles criam para organizar, descrever e de outra forma melhorar o acesso ao objeto de informação (CAPLAN, 2003).

Milstead e Feldman (1999) reiteram esta afirmação, em artigo intitulado "Metadados Catalogando sob um outro nome", deixando claro que o nome metadados pode ser novo, mas a prática é antiga, afirmando que "bibliotecários e indexadores têm produzido e padronizado metadados por séculos". Citam, como exemplo, o primeiro formato de intercâmbio de dados criado para a catalogação automatizada, o MARC (*Machine-Readable Cataloging* – Catalogação Legível por Computador), citado em diferentes momentos desta dissertação.

Segundo Gill (1998), metadados é definido simplemente como dados sobre dados de um objeto de informação. A partir desta definição, o autor faz uma analogia entre metadados e a ficha catalográfica, no sentido de mostrar que a relação existente entre o objeto descrito e os metadados é a mesma que existe entre o livro e a ficha catalográfica. Ainda segundo Gill (1998, p. 9):

> [...] a função do catálogo é apresentar descrições estruturadas dos objetos das coleções com a finalidade de facilitar a busca e recuperação de informações e conseqüentemente o uso e gerenciamento da coleção que está sendo descrita. A descrição do objeto no catálogo objetiva retratar suas características principais. Neste sentido, a informação que está armazenada numa base de dados para gerenciamento de uma coleção de um museu, num inventário computadorizado de um depósito para controle de estoque, numa base de dados composta por registros bibliográficos de uma biblioteca ou ainda num único registro de uma coleção de discos de um indivíduo, é conceitualmente a mesma.

Podemos também fazer um paralelo entre metadados e o processo de indexação. Para demonstrar esta relação, tomamos emprestado a definição de metadados de Milstead e Feldman (1999. p. 1): "os metadados descrevem os atributos e conteúdos de um documento original ou trabalho". As autoras, por sua vez, basearam-se na definição de metadados do Projeto *Development of a European Service for Information on Research and Education* **(**DESIRE): "Dados associados com objetos que isentam seus usuários potenciais de ter conhecimento prévio de sua existência e características". A partir deste conceito, são exemplos de metadados: informação bibliográfica padronizada, sumários, termos indexados e resumos como substitutos do material original. Como já estudado, o processo de indexação nada mais é do que a preparação de uma representação do conteúdo do documento: um registro indexado de um documento é uma representação do documento ou seu substituto, segundo um ponto de vista particular (HARTER, 1986).

Estas técnicas de representação do conteúdo, tão importantes para os sistemas de recuperação da informação, são então utilizadas em um novo ambiente mas com o mesmo objetivo, conforme a definição de metadados de Gomes H., (2000, p. 2), "os metadados nada mais são do que a indicação de categorias de metadados para que os *browsers* possam encontrar as informações requeridas pelos usuários", e complementa, referindo-se aos metadados, como "um aspecto novo para uma velha técnica – a catalogação – já agora em outro contexto e forma, mas basicamente com a mesma finalidade".

Portanto, a prática da catalogação, historicamente percebida como uma arte secreta praticada apenas por bibliotecários, curadores de museus e arquivistas, está se tornando uma questão para uma comunidade mais ampla. Ao mesmo tempo em que, indiscutivelmente, muitas lições podem e devem ser aprendidas dos tradicionais curadores de informação, há também um número de novos desafios característicos do ambiente peculiar da *Web* que vão exigir dos profissionais da informação e bibliotecários uma visão renovada e novas soluções. (GILL, 1998)

É importante frisar que metadados não estão apenas relacionados à descrição de recursos, podendo ter outras funções, como veremos mais adiante mas, quando utilizados para descrever ou identificar recursos de informação, enquanto representação do conteúdo do recurso de

informação, podemos como fizemos, tecer analogias com as práticas de catalogação e indexação de documentos.

É interessante também citarmos a análise de Kraemer (2001) sobre a nova concepção de catalogação-na-fonte, prática que visava a redução de esforços na tarefa de produção da ficha para a composição dos catálogos e consistia na elaboração e impressão da ficha catalográfica no verso da folha de rosto do livro. Podemos estabelecer uma analogia da catalogação-na-fonte com a prática de atribuição de metadados no momento da criação do objeto, que está sendo considerada como a prática mais viável para a catalogação dos recursos disponíveis na *Web*.

Segundo, Milstead e Feldman (1999, p. 3) "não há esperança em catalogar o enorme conjunto de páginas *Web* de uma maneira sistemática" e completam dizendo que a utilização de vocabulários controlados e tesauros por indexadores experientes e treinados consumiria muito tempo para a catalogação da *Web*. Ainda assim, citam que existem inúmeros esforços de voluntários de bibliotecas e grandes organizações de bibliotecas e ainda, esforços de especialistas de áreas particulares, em catalogar a *Web*, como é o caso do projeto *Cataloging and Retrieval of Information Over Networks Aplications* (Catriona II).

Sobre classificação na *Internet*, Souza (2000) entende sua importância para o atendimento às necessidades de informação dos usuários/clientes da *Internet*. Utilizada como instrumento de um sistema tradicional de recuperação da informação, ela é ainda mais necessária no ciberespeaço. O documento Projeto RE 1004 (RE) do DESIRE nos apresenta uma lista de *sites* na *Internet* que utilizam sistemas de classificação da Biblioteconomia ou cabeçalhos de assuntos, disponível no *Beyond Bookmarks*.

Os metadados em documentos na *Web* têm a função de especificar as características dos dados que descrevem, a forma com que serão utilizados, exibidos ou mesmo o significado de seu contexto. As várias definições e aplicações dos metadados, além dos tipos e suas características principais serão estudados no capítulo 5.

5 METADADOS

Neste capítulo, procuramos definir metadados, identificar seus tipos, características e funções, além tipos de entidades para descrição e o entendimento do que se constitui um esquema de metadados.

5.1 Definição de Metadados

Embora metadados seja um tópico de grande interesse para a Ciência da Informação e para a Biblioteconomia, o termo é oriundo da Ciência da Computação. O prefixo "meta" quer dizer "sobre alguma coisa", portanto, uma metalinguagem é uma liguagem utilizada para descrever outras linguagens. Analogamente, metadados são conceituados como dados utilizados para descrever outros dados. A primeira vez que este termo apareceu, neste sentido, foi na primeira edição do *Directory Interchange Format Manual* da NASA, em 1988 (CAPLAN, 2003).

Caplan (2003) conta uma curiosidade: o termo METADATA (em caixa alta) foi cunhado por Jack E. Myers no final de década de 1960 como uma marca registrada da *Metadata Company*, fornecedora de *softwares* e serviços para as áreas de Medicina e Saúde. A palavra *Metadata* era utilizada apenas pela companhia para designar seus produtos. O uso genérico da palavra por outras entidades era permitido, representado pelos termos "meta data" ou "meta-data". Apesar disso, atualmente a maioria das iniciativas de metadados utilizam "metadata" por entenderem que a palavra já é de domínio público.

No início da década de 1990, o termo metadados era atribuído à informação necessária para tornar úteis os arquivos do computador para as pessoas, particularmente os conjuntos de dados científicos, geoespaciais e de Ciências Sociais. Uma das primeiras especificações que se auto-denominou metadata foi a *Content Standard for Digital Geospatial Metadata,* versão 1, do *Federal Geographic Data Comittee*, distribuída em 1994. O objetivo deste padrão era ajudar o usuário a determinar a disponibilidade de um conjunto de dados geoespaciais e sua forma para o uso pretendido, além dos meios necessários para acessar o conjunto de dados geoespaciais e assegurar a transferência bem sucedida dos mesmos (CAPLAN, 2003).

Com o surgimento da *Internet* e da *Web*, o termo metadados começou a ser utilizado no contexto da descrição de objetos de informação na rede. No ambiente da biblioteca, o termo passou a integrar o vocabulário da área no ano de 1995, com a criação e promoção do conjunto de elementos de metadados do Dublin Core. Outra curiosidade apontada por Caplan (2003) é que os organizadores do primeiro *Workshop* do Dublin Core eram participantes ativos do W3C,

naquele tempo uma organização recém-criada, mas já preocupada em gerenciar o desenvolvimento da *Web*, igualmente recente. Desta forma, "a iniciativa do Dublin Core funcionou como um agente para a fertilização cruzada de idéias entre a biblioteca e as comunidades *Web* e foi capaz de energizar os bibliotecários com novos conceitos e terminologia" (CAPLAN, 2003, p. 2).

No ambiente virtual, metadados podem ser utilizados para indicar o nome e a natureza do repositório, certificar a autenticidade e o contexto dos conteúdos e fornecer alguns dados que um profissional da informação ofereceria, como uma referência física (GILL, 1998).

Em ambientes menos tradicionais de informação, o termo metadados é utilizado de forma vasta, como sugerem os exemplos fornecidos por Gilliland-Swetland (1998):

- um provedor da *internet* pode utilizar metadados para se referir à informação codificada em *metatags*[23] em uma página HTML, com o objetivo de tornar mais fácil de achar um *site*;
- profissionais que digitalizam imagens podem pensar em metadados como dados colocados por eles no cabeçalho do arquivo digital para registrar informações sobre a imagem, sobre o processo de visualização e sobre os direitos autorais da imagem;
- um arquivista de Ciências Sociais pode utilizar o termo para designar os sistemas e a documentação de pesquisa, necessários para rodar e interpretar uma fita magnética contendo dados de pesquisa brutos; e
- um arquivista de registros eletrônicos pode adotar o termo ao se referir a toda informação contextual, processada e utilizada para identificar e documentar o escopo, autenticidade e integridade de um documento num sistema eletrônico.

Em todas estas diversas interpretações, metadados são utilizados, não somente para identificar e descrever um objeto informacional, mas também com o propósito de documentar o comportamento do objeto, sua função, uso e gerenciamento, assim como sua relação com outros objetos de informação. Para Caplan (2003), não há interpretação errada ou certa acerca de metadados. A partir dos exemplos anteriores, fica claro que metadados são compreendidos de formas diferentes, dependendo da comunidade e do contexto em que são utilizados. Este pensamento também é compartilhado por Kraemer (2001), que entende que as diferentes definições de metadados levam em consideração suas áreas de aplicação, assumindo diferentes níveis de extensão.

---

[23] *Metatags* são tipos de marcações onde atributos são definidos na forma nome="valor", permitindo que a informação do campo possa ser lida pelos *browsers* e pelos mecanismos de busca e alguma ação possa ser executada a partir de sua identificação.

Caplan (2003, p. 3) conceitua metadados, portanto, a partir de sua utilização: "Metadados são utilizados para significar informação estruturada sobre um recurso de informação de qualquer tipo de mídia ou formato". Nesta definição, não importa se a informação estruturada é ou não eletrônica; se o recurso de informação descrito está ou não sob a forma eletrônica; se é acessível por rede ou disponível pela *Internet*; se é direcionado para o consumo humano ou para o uso da máquina. Contudo, há duas restrições: a informação deve ser estruturada, isto é, não pode ser acumulada aleatoriamente ou representada por um conjunto de elementos de dados que não façam parte de um esquema de metadados. A segunda restrição é que os metadados devem descrever um recurso de informação (CAPLAN, 2003).

De todas as discussões, a que nos parece mais importante é a que está preocupada em entender o que os metadados podem realizar, isto é suas várias aplicações. Neste sentido, um bom exemplo é a definição dada pelo Instituto de Pesquisa Getty em seu Glossário, onde metadados é definido como "dados associados a sistema de informação e a objeto de informação com os seguintes propósitos: descrição, administração, requisitos legais, funcionalidade técnica, uso e preservação".

Outro exemplo é a definição do *U. K. Office for Library and Information Networking* (UKOLN) que se refere a metadados como "dados estruturados sobre recursos digitais (ou não) que podem ser utilizados para dar suporte a um amplo espectro de operações. Estas podem incluir, por exemplo, descrição e descoberta de recursos de informação, seu gerenciamento (incluindo gerenciamento de direitos autorais) e preservação a longo prazo".

5.2 Tipos, características e funções de metadados

Como vimos, todas as concepções sobre metadados são importantes, mas para melhor entendê-las, Gilliland-Swetland (1998) apresenta 05 categorias de metadados: Administrativo, Descritivo, Preservação, Técnico e Uso. O Quadro 2 define cada um destes tipos de metadados e fornece exemplos das funções comuns que cada uma desempenha num sistema de informação digital, que também será abordado.

Quadro 2 - Diferentes tipos de metadados e suas funções

Tipo	Definição	Exemplos
Administrativo	Metadados usados no gerenciamento e administração de recursos de informação.	• Informação sobre aquisição. • Rastreamento da reprodução e dos direitos. • Documentação sobre requisitos de acesso legal. • Informação sobre localização. • Critérios de seleção para digitalização. • Controle de versões.
Descritivo	Metadados usados para descrever ou identificar recursos de informação.	• Registros de catalogação. • Guia de Arquivo.[24] • Índices especializados. • Relações de *hiperlinks* entre recursos. • Anotações feitas por usuários.
Preservação	Metadados usados no gerenciamento da preservação de recursos de informação.	• Documentação sobre a condição física dos recursos. • Documentação sobre ações tomadas para preservar versões físicas e digitais de recursos como, por exemplo, atualização e migração de dados.
Técnico	Metadados usados para retratar o funcionamento de um sistema ou comportamento dos metadados.	• Documentação de *hardware* e *software*. • Informação sobre digitalização, ex: formatos, taxas de compressão, rotinas de *scaling*. • Rastreamento dos tempos de resposta do sistema. • Autenticação e dados de segurança, ex: chaves de *encryption*.
Uso	Metadados usados para mapear o nível e tipo de uso dos recursos de informação.	• Registros de exibição. • Rastreamento do uso e de usuários. • Informação sobre múltiplas versões e reutilização de conteúdo.

Fonte: GILLILAND-SWETLAND, Anne J. Defining Metadata. In: Introduction to Metadata: Pathways to Digital Information. California, 1998, p. 3.

Ainda segundo Gilliland-Swetland (1998), além dos diferentes tipos e funções acima descritos, os metadados também possuem caraterísticas diferentes. O Quadro 3 indica alguns dos principais atributos dos metadados, fornecendo também exemplos ilustrativos.

[24] Traduzimos f*inding aids* como guia. Segundo o *Online Dictionary for Information Science* (ODLIS) *finding aids* é: "Um guia, inventário, índice, registro, calendário, lista ou outro sistema, publicado ou não, para recuperação de materiais arquivísticos de fonte primária que descreve cada item de forma mais detalhada do que a fornecida por um registro catalográfico de biblioteca. *Finding aids* também existe em formatos não-impressos (ASCII, HTML, etc.)".

Quadro 3 - Atributos e características de metadados

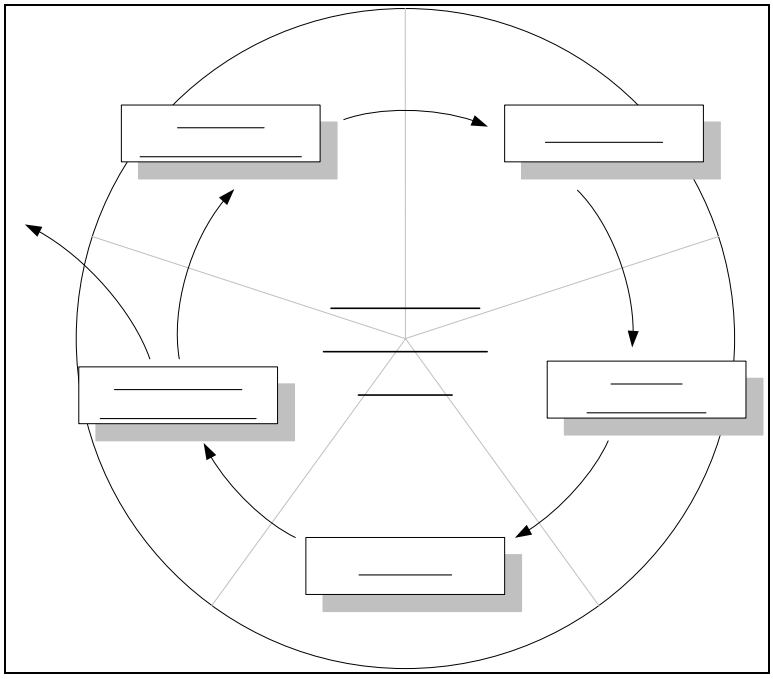
Atributo	Características	Exemplos
Fontes	*Metadados internos* gerados por um agente criador para um objeto de informação no momento de sua criação ou digitalização.	• Nomes de arquivos e informação de cabeçalho. • Estruturas de diretório. • Formato de arquivo e esquema de compressão.
	Metadados externos relacionados a um objeto de informação, criados a *posteriori*, com freqüência por alguém que não é o criador original.	• Registros de catalogação. • Direitos autorais e outras informações de cunho legal.
Método de criação	*Metadados automáticos* gerados por um computador.	• Índices de palavras-chaves. • *Logs* de transações do usuário.
	Metadados manuais criados por pessoas.	• Substitutos descritivos, tais como os registros de catalogação e os metadados Dublin Core.
Natureza	*Metadados não-profissionais* criados por pessoas que não são nem especialistas no assunto nem especialistas de informação, usualmente os criadores originais de um objeto de informação.	• *Metatags* criados para uma página *Web* pessoal. • Sistemas de arquivamento pessoais.
	Metadados profissionais criados ou por um especialista no assunto ou por especialistas de informação, usualmente não sendo o criador original do objeto de informação.	• Cabeçalho de assuntos especializados. • Registros MARC. • Guia de Arquivo.
Status	*Metadados estáticos* que nunca mudam a partir do momento em que foram criados.	• Título, proveniência e dados de criação de um recurso de informação.
	Metadados dinâmicos que podem mudar com o uso/manipulação de objetos de informação.	• Estrutura de um diretório. • *Logs* de transações de usuário. • Resolução de imagens.
	Metadados de longa duração necessários para assegurar que o objeto de informação continue a ser acessível e passível de utilização.	• Informação sobre processamento e formato técnico. • Informação sobre direitos autorais.
	Metadados de curta duração, especialmente de uma natureza transacional.	• Documentação referente ao gerenciamento da preservação.
Estrutura	*Metadados estruturados* que obedecem a uma estrutura previsível, padronizada ou não.	• MARC. • TEI e EAD. • Formatos de bases de dados locais.
	Metadados não-estruturados que obedecem a uma estrutura.	• Campos de notas e anotações não estruturadas.
Semântica	*Metadados controlados* que obedecem a um vocabulário padronizado ou a uma forma de autoridade.	• AAT. • ULAN. • AACR2.
	Metadados não-controlados que obedecem a um vocabulário padronizado ou a uma forma de autoridade.	• Notas de texto livres. • *Metatags* HTML.
Nível	*Metadados de coleção* relacionados às coleções de objetos de informação.	• Registro em nível de coleção, por exemplo, registro MARC ou guia de arquivo. • Índices especializados.
	Metadados individuais relacionados a objetos de informação individuais, usualmente contidos dentro de coleções.	• Legendas transcritas de imagens e datas. • Informação sobre formato.

Fonte: GILLILAND-SWETLAND, Anne J. Defining Metadata. In: Introduction to Metadata: Pathways to Digital Information. California, 1998, p. 4.

Para entendermos como funcionam os metadados, achamos importante reproduzir a Figura 2, também de autoria de Gilliland-Swetland (1998), pois é muito esclarecedora a respeito do papel desempenhado pelos metadados nos diferentes estágios da vida de um objeto de informação num ambiente virtual.

Segundo Gilliland-Swetland (1998), a criação e gerenciamento de metadados se tornou um *mix* complexo de processos manuais e automáticos e de camadas criadas por muitos indivíduos e funções em momentos diferentes na vida de um objeto de informação. Pelo que podemos ver na Figura 2, de uma fase para outra, os objetos adquirem camadas de metadados que podem estar associadas com os objetos de diferentes formas. Os metadados podem estar contidos dentro do próprio objeto de informação como, por exemplo, no cabeçalho de um arquivo de imagem. Metadados podem estar anexados ao objeto de informação através de apontadores bi-direccionais ou *hiperlinks*. As relações entre metadados e objetos de informação e entre diferentes aspectos de metadados podem também ser documentados em um *registry*, como veremos mais adiante.

Figura 2 - Ciclo de vida dos objetos contidos num sistema de informação digital



Fonte: GILLILAND-SWETLAND, Anne J. Defining Metadata. In: Introduction to Metadata: Pathways to Digital Information. California, 1998, p. 4.

A seguir, a descrição de cada uma das fases do ciclo de vida de um objeto de informação, de acordo com Gilliland-Swetland (1998, p. 5):

- **Criação e Multiplas Versões:** Objetos entram o sistema de informação digital, criados em forma digital ou convertidos na forma digital. Múltiplas versões do

> mesmo objeto podem ser criadas para preservação, pesquisa, disseminação ou até para desenvolvimento de produtos. Alguns metadados administrativos e descritivos podem ser incluídos pelo criador.
>
> - **Organização:** Objetos são automaticamente ou manualmente organizados na estrutura de um sistema de informação digital e metadados adicionais podem ser criados através dos processos de registro, catalogação e indexação.
>
> - **Busca e Recuperação**: Objetos armazenados e distribuídos são passíveis de busca e recuperação pelos usuários. Um sistema de computador cria metadados que rastreiam algoritmos de recuperação, transações de usuário e a eficácia do sistema no armazenamento e recuperação.
>
> - **Utilização**: Objetos recuperados são utilizados, reproduzidos e modificados. Metadados referentes às anotações do usuário, mapeamentos dos direitos autorais e controle de versões podem ser criados.
>
> - **Preservação e Disponibilização**: Os objetos de informação sofrem processos como revigoração, migração e checagem da integridade para assegurar sua contínua disponibilidade. Objetos de informação que são inativos ou não mais necessários podem ser descartados. Os metadados podem documentar tanto as atividades de preservação quanto de disponibilização.

Caplan (2003) e Gilliland-Swetland (1998) destacaram alguns aspectos que devem ser considerados na definição e utilização dos metadados, entendidos pelas autoras como "mitos" e que achamos importante aqui reproduzir:

- *Metadados não se referem apenas à descrição de recursos*, podem ser utilizados para administração, acesso, preservação e uso de coleções como o fazem os museus virtuais, bibliotecas e arquivos digitais.
- *Metadados não precisam ser eletrônicos.* Se isto não fosse verdade, implicaria dizer que um registro MARC é metadados, enquanto uma ficha catalográfica, ainda não convertida para este formato, não é metadados. Até mesmo dentro da própria comunidade de bibliotecários, nota-se uma inconsistência: alguns se referem a metadados apenas para a descrição de recursos eletrônicos, enquanto outros se referem a metadados como a descrição de quaisquer recursos, eletrônicos ou não. Embora o conceito mais restrito seja o mais próximo ao conceito original da Ciência da Computação, é certamente mais lógico pensar em metadados como descrição de todos os tipos de recursos de informação.
- *Metadados provêm de uma variedade de fontes*: podem ser fornecidos por humanos (o criador, o profissional da informação ou o usuário) ou criados automaticamente por um computador ou ainda inferidos através de sua relação com outro recurso, tal como o *hyperlink*.
- *Metadados podem ser acrescidos durante o tempo de vida de um objeto de informação:* metadados podem ser criados, modificados ou até mesmo descartados durante a vida de um recurso.

5.3 Tipos de entidades para descrição

Nesta seção, procuramos apresentar os tipos de entidades que os metadados descrevem, pois os metadados podem ser utilizados para descrever muitos tipos ou níveis de entidades, de conceitos abstratos a objetos físicos. Achamos essa discussão importante, pois na definição de um esquema ou elemento de metadados é fundamental especificar os tipos de entidades aos quais se referem.

Para a definição dos tipos de entidades, baseamos nossa análise no estudo de Caplan (2003) sobre o modelo descrito no *IFLA Functional Requirements for Bibliographic Records* (FRBR), que estabelece quatro níveis de entidades: trabalho (*work*), expressão (*expression*), manifestação (*manifestation*) e item (*item*).

Um **trabalho** é um conceito abstrato definido como uma criação artística ou intelectual distinta. Um trabalho pode ter muitas **expressões**, incluindo diferentes edições, traduções, condensações e arranjos. Por exemplo, Otelo de Shakespeare é um trabalho, mas uma edição particular de Otelo é uma expressão. Contudo, uma modificação que introduz novos aspectos intelectuais e artísticos é considerada um novo trabalho. Neste sentido, a ópera Otelo de Verdi deve ser considerada como um outro trabalho, porque possui seu próprio conjunto de expressões na forma de partituras, livretos e performances. (CAPLAN, 2003)

Uma **manifestação** é definida como a personificação física da expressão de um trabalho ou todas as cópias de uma expressão produzida na mesma mídia e forma física. Uma performance da ópera Otelo de Verdi, por exemplo, pode ser gravada em filme, DVD, VHS, CD e vários formatos de fita cassete. Cada uma delas constitui separadamente uma manifestação. (CAPLAN, 2003)

A última entidade neste modelo é o **item**, definido como um exemplar único de uma manifestação, um único objeto físico, ou um conjunto de objetos físicos (por exemplo, uma monografia em dois volumes ou uma gravação em cd duplo). (CAPLAN, 2003)

É importante observar que o modelo da FRBR não contempla todas as entidades e que a maioria dos esquemas de metadados possuem elementos que pertencem a mais de uma entidade da FRBR. O importante é frisar que um esquema de metadados deve dispor de um modelo explícito que descreva os tipos de entidades, considerando também suas possíveis relações.

5.4 Esquema de metadados

Antes de analisarmos o esquema Dublin Core, precisamos, em primeiro lugar, entender o que se constitui um esquema de metadados e também analisar os três aspectos que lhe são próprios: semântica, regras de conteúdo e sintaxe.

Um esquema (scheme[25]) de metadados é um conjunto de elementos de metadados e regras para seu uso, definidos para um propósito em particular e, segundo Caplan (2003) pode apresentar três aspectos: semântica, regras de conteúdo e sintaxe.

A semântica refere-se ao significado dos itens de metadados (elementos de metadados). Um esquema de metadados especifica os elementos de metadados do esquema, atribuindo-lhes um nome e uma definição. O esquema deverá também indicar se o elemento é obrigatório ou opcional, ou se pode ou não ser repetido (CAPLAN, 2003).

As regras de conteúdo especificam como os valores atribuídos aos elementos de metadados são selecionados e representados. Por exemplo, a semântica de um esquema de metadados define o elemento denominado "autor", enquanto as regras de conteúdo especificam informações, tais como, que agentes são qualificados como autores e como o nome do autor deve ser registrado (a sua representação) (CAPLAN, 2003). As regras de conteúdo normalmente determinam o uso de instrumentos como o tesauro ou como um esquema de classificação, já analisados anteriormente na seção 3.3., enquanto instrumentos utilizados para recuperação da informação, ferramentas tão relevantes aos sistemas de recuperação da informação tradicionais, agora ainda mais importantes para a recuperação da informação no ambiente virtual.

A sintaxe de um esquema representa como os elementos são codificados em linguagem legível pelo computador. Em termos gerais, os sistemas de processamento designados para buscar, mostrar ou atuar sobre os metadados podem ter formatos de armazenamento interno bem diferentes dos formatos de metadados (CAPLAN, 2003). Uma sintaxe específica de um esquema serve mais para prover um formato de intercâmbio comum para troca de metadados entre as partes do que para prescrever como os dados são armazenados num sistema local, assim como ocorreu no mundo das bibliotecas com o formato MARC, analisado anteriormente. Segundo Caplan (2003), a sintaxe de um esquema de metadados pode ser chamada de formato de comunicação (*communication format*), formato de intercâmbio (*exchange format*), sintaxe de transporte (*transport syntax*) ou sintaxe de transmissão (*transmission syntax*).

---

[25] Caplan (2003) faz uma distinção entre *scheme* e *schema*. O termo *schema* possui um outro significado relacionado à tecnologia de bases de dados de computador, sendo definido como a organização formal ou estrutura de uma base de dados, ou utilizado em referência ao XML. No Brasil, normalmente *scheme* é traduzido como padrão.

A semântica, as regras de conteúdo e a sintaxe são independentes, mas aspectos relacionados entre si. Na prática, qualquer esquema em particular pode conter, misturar ou omitir estes componentes em qualquer combinação. Por exemplo, alguns esquemas de metadados são definidos como estruturas SGML ou XML, em que a semântica está intrincadamente emaranhada com a sintaxe. Outros esquemas de metadados não especificam nenhuma sintaxe ou, ainda, oferecem aos implementadores múltiplas sintaxes para sua escolha. Alguns esquemas não contêm regras de conteúdo ou se referem a regras de conteúdo externas e podem ser desenhados para permitir o uso de quaisquer regras de conteúdo, desde que o conjunto de regras seja especificado (CAPLAN, 2003).

5.4.1 Sintaxe de Metadados[26]

Nesta seção, analisamos alguns dos formatos utilizados para representar metadados em forma legível por computador. Em alguns casos, os metadados são armazenados e processados em sistemas locais nestes formatos. Mas, em termos gerais, os metadados são armazenados em bases de dados locais mas trocados com outros sistemas utilizando estes formatos como sintaxes de transporte. Neste caso, o sistema local precisará importar ou exportar metadados em um ou mais desses formatos.

As seguintes sintaxes serão analisadas: MARC, SGML, HTML, XML e RDF.

5.4.1.1 MARC

A sintaxe mais utilizada nas bibliotecas é o MARC. É importante destacar, segundo Kraemer (2001), que o esquema MARC é composto por um conjunto de regras e especificações de formato utilizadas na catalogação tradicional das bibliotecas, que inclui a *International Standard Bibliographic Description* (ISBD), as *Anglo-American Cataloguing Rules* (AACR), as especificações do MARC21[27] e um número de documentos de referência. A AACR é publicada pelas associações de bibliotecas americana, canadense e inglesa e está disponível na forma impressa e em CD-ROM (WEBER, 2002).

Além das especificações do *MARC21 Format for Bibliographic Data*, a sintaxe de transporte é constituída também pelo formato de transmissão de dados especificado pela ANSI/NISO Standard Z39.2. O padrão Z39.2 define um formato para transmissão de dados, que consiste em três partes: **cabeçalho**, **diretório** e número variável de **campos** (cada campo pode ser um **campo de controle** ou um **campo de dados**) (CAPLAN, 2003).

---

[26] A sintaxe de metadados é denominada por Kraemer (2001) "linguagem de marcação para descrição de metadados", ou "comandos de marcação", segundo Souza e Alavarenga (2004).

O **cabeçalho** contém 24 *bytes*, agrupados em nove elementos de dados e cada elemento pode ser um código ou um contador (CAPLAN, 2003).

O **diretório** contém um número de entradas igual ao número de campos de dados que estão sendo transmitidos. Cada entrada possui 12 *bytes* que estão agrupados em três elementos: nome ou *tag* (três *bytes*), cumprimento e posição de início do campo de dados, ao qual a entrada faz referência (CAPLAN, 2003).

Um **campo de controle** contém um número pré-definido de *bytes* e, da mesma forma que o cabeçalho, está segmentado em elementos de dados com significados específicos (CAPLAN, 2003).

O **campo de dados** começa com dois indicadores – consistindo de um *byte* cada um, seguidoS de dados textuais subdivididos em sub-campos e terminando com um *byte* finalizador**.** Os sub-campos são delimitados por um *byte* conhecido como delimitador de sub-campo (uma barra vertical ou um símbolo de dólar), seguido de um código de um *byte* que indica o tipo de sub-campo (CAPLAN, 2003).

5.4.1.2 SGML

A linguagem *Standard Generalized Markup Language* (SGML) é um padrão internacional (*ISO 8879:1986 Information processing – Text and office systems*), formalmente definida como uma metalinguagem ou uma linguagem para descrição de outras linguagens. Ela especifica regras genéricas de sintaxe para a codificação dos documentos, mas não especifica nenhum conjunto particular de *tags*. Ao invés disso, oferece os meios para que a pessoa possa definir seu próprio conjunto de *tags* e regras de uso. Isso é feito através da criação de um "Document Type Definition" (DTD). Um DTD, por exemplo, poderia ser chamada "HTML" e especificar que o conjunto de *tags* permitido são: <TITLE>, <META>, <LINK>, <HEAD>, <BODY> e <P> e que os tags <TITLE>, <META> e <LINK> deverão aparecer dentro da <HEAD>, enquanto <P> deverá aparecer dentro da <BODY> (CAPLAN, 2003). Desta forma, podemos entender que a linguagem HTML é na verdade um DTD específico da SGML. (SOUZA e ALVARENGA, 2004). A linguagem HTML será analisada na seção 5.4.1.3.

---

[27] O MARC21 surgiu em 1988 como parte de um esforço para harmonizar os formatos MARC americano, canadense e britânico.

As marcações da SGML codificam os elementos de dados entre as *tags* inicial e final e outros elementos de dados como valores para atributos, os quais seguem depois do nome da *tag* inicial. Por exemplo, a tag <META> possui o atributo "NAME", cujo valor aparece depois do caractere "=":

<META NAME="title" ...>

Os atributos podem ser definidos como opcionais ou obrigatórios. Uma lista de valores pode ser especificada contendo os atributos permitidos. Um elemento da SGML pode ser definido para conter dados textuais e um ou mais atributos, como também conter somento texto ou somente atributos. Além disso, um elemento pode conter outros elementos. Por exemplo: o elemento <DATE> pode conter os elementos <MONTH>, <DAY> e <YEAR>. Outros elementos podem ser definidos para não conter nem texto nem outros elementos. Por exemplo, a tag <lb> indica uma quebra de linha (KRAEMER, 2001).

A SGML é considerada uma boa linguagem de codificação para metadados por vários motivos: permite a utilização de dados textuais de comprimento variável; permite definir um ilimitado número de elementos (*tags* e atributos), cujos nomes são representativos de seu conteúdo e possibilita expressar as relações hierárquicas encontradas dentro de coleções e entre trabalhos, expressões, manifestações e itens. Além disso, é uma linguagem flexível para definir metadados, pois um elemento SGML pode conter outros elementos. Apesar destas vantagens, há uma desvantagem significante: a SGML é uma linguagem difícil para ser processada pelos programas. Por este motivo, são poucos *softwares* que suportam a criação, armazenamento e modificação da linguagem SGML (CAPLAN, 2003).

Um exemplo da definição de uma *tag* num documento DTD pode ser visto na Figura 3.

Figura 3 - Exemplo da definição de uma *tag* num documento DTD

<DIV> Text Division

Description:
A generic element that designates a major section of text within <frontmatter>. Examples of these divisions include a title age, preface, acknowledgments, or instructions for using a finding aind. Use the <HEAD> element to identify the <DIV>'s purpose.

May contain:
Address, blockquote, chronlist, div, head, list, note, p, table

May occur within:
div, frontmatter

Attributes:

ALTRENDER	#IMPLIED, CDATA
AUDIENCE	#IMPLIED, external, internal
ID	#IMPLIED, ID

Fonte: CAPLAN, Priscilla. Metadata fundamentals for all librarians. Chicago: American Library Association, 2003, p.19 .

5.4.1.3 HTML

A linguagem *Hiper Text Mark-up Language* (HTML) é uma aplicação especial e limitada da sua originária, a SGML, usada para codificar documentos a serem disponibilizados por meio de servidores de rede e utilizados por meio de navegadores na *Web* (KRAEMER, 2001). Dentre as vantagens desta linguagem, podemos citar a simplicidade, caráter genérico e seu alto grau de utilização e implantação. Por outro lado, Souza e Alvarenga (2004, p. 5) caracterizam a estrutura do HTML como "rígida, não existindo a possibilidade de adição de novos comandos de marcação (*tags*), sem que haja uma redefinição do DTD da linguagem e conseqüente atualização dos navegadores para que interpretem estas novas *tags*".

A HTML utiliza marcações ou *tags* pré-definidas, em meio ao texto, para delimitá-lo. A maioria das *tags* trabalha em pares com uma *tag* abrindo e outra fechando, sendo ambas iguais com exceção do caractere "/", que inicia sempre a *tag* de fechamento, por exemplo:
<TITLE> Título do documento</TITLE>.

Um documento HTML começa com uma *tag* <HTML> e termina com uma *tag* </HTML>. No interior destas *tags*, o documento é dividido em duas outras seções: <HEAD> e <BODY>. Dentro da seção <HEAD>, aparecem as *tags* <TITLE> e <META>. O conteúdo real da página *Web* aparecerá na seção <BODY> (CAPLAN, 2003).

Segundo Caplan (2003), os metadados podem ser inseridos no documento HTML, utilizando a *tag* <META>. A forma mais utilizada é:
<META NAME="text string1" CONTENT="text string2">

O nome do elemento de metadados corresponde ao "text string1", enquanto o valor do elemento é representado no "text string2", como no exemplo a seguir:
<META NAME="author" CONTENT="Rosa, Guimarães">

Para o atributo NAME, pode ser utilizado qualquer rótulo, que só será útil se reconhecido pelos mecanismos de busca para recuperação. Muitos mecanismos de busca na *Internet* reconhecem pelo menos alguns elementos do Dublin Core e qualquer mecanismo de busca pode ser programado para reconhecer elementos de qualquer esquema. Uma prática recomendada é explicitar o esquema que está sendo utilizado para a especificação do elemento - na forma de um prefixo - e utilizar a *tag* <LINK> para associar o prefixo à definição do esquema disponível na *Web*, como por exemplo:
<META NAME="DC.Creator" CONTENT="Rosa, Guimarães">
<LINK REL="schema.DC" HREF="http://purl.org/DC/elements/1.0/">

Um exemplo completo de metadados embebidos num documento HTML pode ser visto na Figura 4.

Figura 4 - Exemplo completo de metadados embebidos num documento HTML

```
<HTML>
<HEAD>
< TITLE >Weather Report for Monday</TITLE>
<META NAME="DC.Title" CONTENT=" Weather Report for Monday">
<META NAME="DC.Creator" CONTENT="National Weather Service">
<META NAME="DC.Date" CONTENT=" 2001-12-01">
<LINK REL="schema.DC" HREF="http://purl.org/DC/elements/1.0/">
</HEAD>
<BODY>
<P>Warmer and slightly cloudy with a 20% chance of afternoon thunderstorms</P>
</BODY>
</HTML>
```

Fonte: CAPLAN, Priscilla. Metadata fundamentals for all librarians. Chicago: American Library Association, 2003, p. 16.

5.4.1.4 XML

A partir das necessidades de uma linguagem que descrevesse o conteúdo semântico e os significados contextuais, além da estrutura e forma de exibição de documentos, foi criado o *Extensible Markup Language* (XML) (SOUZA e ALVARENGA, 2004). Como já abordado, o XML é uma recomendação formal do W3C.

Tanto a XML como a HTML são originárias da linguagem SGML. A linguagem XML pode ser pensada como um subconjunto da SGML, desenhada com regras mais rigorosas, menos características e opções, tudo para que o processamento seja mais fácil. Por exemplo, na linguagem SGML, as *tags* finais podem ser omitidas sob certas circunstâncias e os valores dos atributos podem ou não estar entre aspas. Na linguagem XML, ao contrário, se um elemento possui uma *tag* final, seu uso é obrigatório e um valor de atributo deve sempre aparecer entre aspas (CAPLAN, 2003).

A XML foi desenvolvida em parte para resolver as limitações da linguagem HTML: "enquanto a HTML tem como objetivo controlar a forma com que os dados serão exibidos, a XML se concentra na descrição dos dados que o documento contém. Além disso, a XML é flexível no sentido de que podem ser acrescentadas novas *tags* a medida em que forem necessárias, bastanto para isso que estejam descritas em um DDT específico (SOUZA e ALVARENGA, 2004, p. 5).

Caplan (2003) confirma a importância desta linguagem quando se refere ao fato de que os esquemas de metadados estão sendo definidos como XML DTDs e que é bastante provável que os futuros esquemas de metadados sejam definidos usando esquemas XML, ao invés de DTDs.

5.4.1.5 RDF

O *Resource Description Framework* (RDF) é um modelo de dados para representar recursos, suas propriedades e os valores destas propriedades e, em teoria, este modelo de dados pode ser representado em qualquer sintaxe. Segundo Caplan (2003), quando se pensa em RDF, geralmente pensamos em sua representação em XML.

O conceito fundamental do RDF é a noção de *namespace*. Um *namespace* é definido por Souza e Alvarenga (2004, p. 8) como "um vocabulário controlado que identifica um conjunto de conceitos, de forma única para que não haja ambigüidade na sua interpretação. Os *namespaces* XML são conjuntos de tipos de elementos e atributos possíveis para cada tipo".

Cada elemento de metadados numa descrição RDF é precedido por um rótulo associando o elemento a um *namespace* em particular. A partir da utilização do *namespace*, dois objetivos são atingidos: em primeiro lugar, o nome do elemento de metadados é associado com uma forma de obter sua definição e em segundo lugar, os elementos de vários esquemas de metadados podem ser usados juntos sem ambigüidade para descrever um único recurso (CAPLAN, 2003).

Alguns benefícios do padrão RDF são apontados por Souza e Alvarenga (2004, p. 8):

- prover um ambiente consistente para a publicação e utilização de metadados na *Web*, utilizando a infra-estrutura do XML;
- prover uma sintaxe padronizada para a descrição dos recursos e propriedades dos documentos na *Web*;
- permitir que as aplicações possam agir de forma inteligente e automatizada sobre as informações publicadas na *Web*, uma vez que seus significados são mais facilmente intelegíveis.

Um exemplo simples de representação em RDF pode ser visto na Figura 5:

Figura 5 - Exemplo de representação em RDF

```
<?xml version="1.0"?>
<rdf:RDF
        xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntas-ns#"
        xmlns:dc="http://purl.org/dc/elements/1.1/">
<rdf:Description about="http://[URL of weather report page]">
        <dc:title>Weather Report for Monday</dc:title>
        <dc:creator>National Weather Service</dc:creator>
        <dc:date>2001-12-01</dc:date>
</rdf:Description>
</rdf:RDF>
```

Fonte: CAPLAN, Priscilla. Metadata fundamentals for all librarians. Chicago: American Library Association, 2003, p. 21.

5.5 Interoperabilidade

No ambiente da rede, há muitos tipos de interoperabilidade. Podemos pensar em interoperabilidade como o compartilhamento de um protocolo comum de comunicação entre duas aplicações, por exemplo, ou como a possibilidade de um cliente interagir com muitos servidores ou, até mesmo, a utilização de dados em diferentes contextos (CAPLAN, 2003).

Segundo Gomes, S. (2002, p. 77), "interoperabilidade é um termo amplo que compreende questões relacionadas à possibilidade de bases de dados e outros recursos distribuídos trabalharem juntos, oferecendo ao usuário a capacidade de acessá-las mediante a busca "atravessada"[*cross-search*] ou mediante navegação [*cross-browse*], a partir de uma única interface" (p. 77). A autora adverte que a interoperabilidade "requer concordância em três níveis: técnico, organizacional e de conteúdo" (ARMS *apud* GOMES, S., 2002, p. 77) , enfatizando que "a questão transcende abordagens que apenas privilegiem o aspecto da tecnologia para viabilizar as operações, em detrimento dos demais" (GOMES, S., 2002, p. 77).

Caplan (2003) se refere aos tradicionais catálogos unificados de bibliotecas baseados no formato MARC como o *WorlCat*, da OCLC, em nível internacional, já citado anteriormente, como um exemplo de interoperabilidade que permite a busca numa única base de dados central de metadados de vários recursos. Neste caso, é utilizado apenas um único sistema para fazer a busca e recuperar informações.

Há uma outra forma de conseguir os mesmos resultados onde os registros de metadados são armazenados em várias bases de dados distribuídas. O que possibilita esta busca atravessada [*cross-search*] é o protocolo internacional Z39.50 que é "um protocolo de comunicação entre computadores desenhado para permitir pesquisa e recuperação de informação - documentos com textos completos, dados bibliográficos, imagens, multimeios - em redes de computadores

distribuídos. Baseado em arquitetura cliente/servidor e operando sobre a rede *Internet*, o protocolo permite que a pesquisa seja realizada em vários sistemas de informação distribuídos por meio de única interface de busca". (ROSETTO, 1997, p. 1)

Alguns autores apregoam que o modelo de interoperabilidade baseado no protocolo Z39.50 é mais eficiente do que o modelo em que a busca é realizada numa única base de dados central. Mas Caplan (2003) aponta também desvantagens: o "cliente" Z39.50 só pode falar com um "servidor" Z39.50 e nem todos os serviços de informação *online* possuem servidores Z39.50.

Quando falamos de interoperabilidade no contexto dos metadados, falamos da habilidade de realizar uma busca entre diferentes conjuntos de metadados e obter resultados significantes. Neste caso, os metadados podem ter sido criados de acordo com o mesmo esquema, mas por diferentes indivíduos ou organizações, ou podem ter sido criados a partir de diversos esquemas. (CAPLAN, 2003). Para facilitar a interoperabilidade entre diferentes esquemas de metadados, vamos abordar, mais especificamente, as *crosswalks* e os r*egistries* nas seções seguintes. A maior parte de nossa explanação se baseia no trabalho de Caplan (2003) que, dentre outros artigos, foi o mais sistematizado e nos possibilitou uma visão clara do assunto.

5.5.1 *Crosswalks*

A interoperabilidade entre diferentes esquemas de metadados é facilitada pelo uso de *crosswalks* que tem a função de mapear os elementos de um esquema para outro. (CAPLAN, 2003) Podemos citar, como exemplo de *crosswalk*, o mapeamento dos elementos do Dublin Core para os elementos do MARC, feito pela *Library of Congress,* que especifica, por exemplo, que o elemento "Colaborador" do Dublin Core não-qualificado pode ser mapeado para o campo 720 do MARC. A Figura 6 mostra um exemplo de *crosswalk* entre o Dublin Core/MARC e GILS.

Figura 6 - Exemplo de *crosswalk* entre o Dublin Core/MARC e GILS

Creator
An entity primarily responsible for making the content of the resource.

MARC 21:

Unqualified:

- 720 ##$a (Added Entry--Uncontrolled Name/Name) with $e=author

Qualified:

- Personal: 700 1#$a (Added Entry--Personal Name) with $e=author
- Corporate: 710 2#$a (Added Entry--Corporate Name) with $e=author
- Conference: 711 2#$a (Added Entry--Conference Name) with $e=author
- Role: 720 ##$e (Added Entry--Uncontrolled Name/Relator term
- Role (Personal): 700 1#$e (Added Entry--Personal Name/Relator term)
- Role (Corporate): 710 2$e (Added Entry--Corporate Name/Relator term)

Note: The above qualifiers have not been approved by DCMI.

GILS:

- Originator

Fonte: Dublin Core/MARC/GILS Crosswalk. Disponível em http://lcweb.loc.gov/marc/dccross.html. Acesso em: 26/07/04.

Crosswalks foram feitas para a maioria dos grandes esquemas para descrever recursos na *Web*. A *Library of Congress* também mantém mapeamentos do MARC21 para vários esquemas e de vários esquemas para o MARC21. Dentre estes vários esquemas, podemos citar, além do Dublin Core e do GILS, o FGDC Content Standards for Geospatial Metadata e o ONIX. (CAPLAN, 2003)

As *crosswalks* podem ser utilizadas como especificações básicas para a conversão de um esquema de metadados para outro esquema para possibilitar a troca de registros ou pelos mecanismos de busca, para varrer campos com o mesmo conteúdo ou similar, em diferentes bases de dados e, ainda, podem ser utilizadas para ajudar os profissionais de informação no entendimento de novos esquemas de metadados.

5.5.2 Registries

Os *registries* podem ser entendidos como ferramentas utilizadas para registrar informações de autoridade sobre os elementos de metadados de inúmeras fontes (CAPLAN, 2003). Ao falar sobre o Dublin Core, Weibel (2000, p. 6) atesta a importância dos *registries* para a interoperabilidade:

> Há muitas aplicações do Dublin Core (e de metadados em geral) que alguém possa facilmente acompanhar. Os *registries* das aplicações, contendo as definições de seus elementos e da semântica, poderiam economizar bastante tempo para novos implementadores, gerando uma crescente consciência e cooperação, e, acima de tudo,

> apoio para uma interoperabilidade mais ampla entre aplicações e coleções de metadados.

Através do registro de nomes, definições e propriedades dos elementos de metadados, os *registries* facilitam a identificação, reutilização e interoperabilidade entre os elementos. A partir do crescente número de esquemas de metadados, os *registries* estão assumindo um papel preponderante como ferramentas de gerenciamento de dados (CAPLAN, 2003).

O Projeto DESIRE entende os *registries* importantes para:

- as pessoas que querem criar metadados de acordo com padrões definidos, para os que querem descobrir se conjuntos de elementos apropriados já existem o propósito pretendido e para aqueles que querem alinhar seus conjuntos de elementos com outros que são utilizados para outros objetivos.

- os *softwares* que querem manipular metadados e necessitam saber sua estrutura e semântica, para as ferramentas que criam metadados e que precisam validar e apresentar uma interface para o usuário e, por último, para as ferramentas de conversão que precisam de referência para o mapeamento de tabelas.

A maioria dos *registries* são baseados num padrão ISO/IEC 11179 *Standard, Specification and Standardization of Data Elements*. Um dos *registries* mais conhecidos baseado neste padrão, é o *Australian Institute of Health and Welfare Knowledgebase* que inclui definições de elementos relacionados à saúde, serviços comunitários e assistência domiciliar. O maior dos Estados Unidos, também baseado neste padrão, é o *Environmental Data Registry* (EDR) da Environmental *Protection Agency* (CAPLAN, 2003).

Caplan (2003) cita ainda três outros exemplos de *registries* que não seguem o padrão *ISO/IEC 11179*. O primeiro é o da *Resource Organization and Subject-based Services* (ROADS) que é um projeto do Programa *Eletronic Libraries* (eLib) do *Joint Information Systems Comittee* (JISC) do Reino Unido, que se constitui, na verdade, numa lista de *templates* (modelos) e dos elementos que contêm. O segundo exemplo é o *registry* do *Dublin Core Metadata Initiative* (DCMI), um projeto de pesquisa da OCLC, como apontamos anteriormente, utilizado como uma ferramenta para utilização dos usuários finais para obter informações sobre os termos do Dublin Core, seu uso e suas relações. O terceiro exemplo é o *registry* do *Development of a European Service for Information on Research and Education* (DESIRE), muitas vezes relacionados nesta dissertação, que gerencia elementos de metadados de múltiplos *namespaces* (esquemas) (CAPLAN, 2003).

6 MAPEANDO METADADOS NO EXTERIOR E NO BRASIL

Este capítulo, embora ainda aborde alguns conceitos da literatura, enfoca os resultados da etapa de análise empírica, conforme os procedimentos metodológicos descritos ao final da introdução.

Na primeira parte analisamos o padrão Dublin Core, apresentamos seus elementos constituintes e fazemos também um contraponto com o formato MARC. Depois, tecemos considerações a respeito do mapeamento dos esquemas no exterior, através de um levantamento de 27 esquemas de metadados utilizados por várias comunidades no exterior para, finalmente, apresentarmos os resultados de nossa pesquisa sobre a utilização de metadados em sistemas de informação no Brasil.

6.1 Análise do padrão internacional Dublin Core

O padrão Dublin Core surgiu de um *Workshop* realizado pela *Online Computer Library Center* (OCLC) e pelo *National Center for Computing Applications* (NCSA), em março de 1995, em Dublin, Ohio, nos Estados Unidos. Vários profissionais de diversos campos de atuação participaram do evento, tais como bibliotecários, cientistas da computação, cientistas da informação, indexadores, museólogos, arquivistas, e outros. (SOUZA, CATARINO e SANTOS, 1997). Neste W*orkshop* foram estabelecidos 13 elementos mínimos para a descrição de recursos e, em setembro de 1996, num outroW*orkshop*, também realizado em Dublin, foram acrescidos dois elementos, totalizando 15 elementos denominados *Dublin Metadata Core Element Set*, conhecido como Dublin Core (HUDGINS et al, 1999).

Segundo Weibel et al (1998), as metas que motivaram o esforço para a crição do Dublin Core foram: a simplicidade de criação e manutenção, semântica facilmente compreendida, conformidade com padrões já existentes e em formação, aplicabilidade e abrangência internacionais, extensibililidade, interoperabilidade entre coleções e sistemas de indexação.

O padrão de metadados Dublin Core foi criado, portanto, com o intuito de possibilitar a descrição padronizada de qualquer tipo de recurso na *Web* e suas características são ratificadas por Souza, Vendrúsculo e Melo (2000) e por Kraemer (2001, p. 40), quando aborda os objetivos da criação deste padrão:

> A concepção de um formato que unifique os dados necessários para descrever, identificar, processar, localizar e recuperar recursos virtuais, beneficiando mantenedores e usuários de sistemas de informação introduzidos em redes, levou profissionais e entidades a estabelecerem padrões mínimos que direcionam aplicações de metadados.

> Dentre estas iniciativas, destaca-se o padrão Dublin Core, o qual está caminhando para assumir um caráter de padrão internacional, uma vez que tem tido ampla aceitação das comunidades virtuais.

O *Dublin Core Metadata Initiative* (DCMI) gerencia o desenvolvimento de especificações oficiais relacionadas ao Dublin Core, mantido por curadores e por um grande número de voluntários.

Todos os elementos do Dublin Core são opcionais, repetitivos e podem ser dispostos em qualquer ordem. Os 15 elementos são divididos nas seguintes categorias de informação, conforme o Quadro 4: Conteúdo (título, assunto, descrição, fonte, idioma, relação, cobertura), Propriedade Intelectual (criador, editor, colaborador, direitos autorais) e Manifestações Físicas[28](data, tipo, formato, identificador) (HUDGINS, 1998).

Quadro 4 - Elementos do Dublin Core por categorias de informação

Conteúdo	**Propriedade intelectual**	**Manifestação física**
Título	Criador	Data
Assunto	Editor	Tipo
Descrição	Colaborador	Formato
Fonte	Direitos Autorais	Identificador
Idioma		
Relação		
Cobertura		

Fonte: WEIBEL, S., KUNZE, J., LAGOZE, C., WOLF, M. Dublin Core Metadata for Resource Discovery. IETF #2413. The Internet Society, September, 1998. Disponível em: http://www.ietf.org/rfc/rfc2413.txt. Acesso em: 28.07.04.

A seguir, no Quadro 5, a descrição pormenorizada de cada um dos elementos, com os seguintes dados: nome do elemento, rótulo, definição e comentário.

[28] O termo "Instantiation" é traduzido por Kraemer (2001) como "Manifestação Física", e traduzido por Weber (2002) por "Representação" (*Representation*).

Quadro 5 - Descrição dos elementos do Dublin Core

Nome do elemento: Título	
Rótulo:	Título
Definição:	Nome dado ao recurso.
Comentário:	Normalmente, o Título será um nome pelo qual o recurso é formalmente conhecido.
Nome do elemento: Criador	
Rótulo:	Criador
Definição:	Entidade principal responsável pelo conteúdo intelectual do recurso.
Comentário:	Exemplos de Criador incluem uma pessoa, uma organização ou um serviço. Normalmente, o nome do Criador deve ser usado para indicar esta entidade.
Nome do elemento: Assunto	
Rótulo:	Assunto ou Palavras-chaves
Definição:	Tópico do conteúdo do recurso.
Comentário:	Normalmente, o Assunto será expresso por palavras-chaves, frases ou códigos de classificação que descrevem o assunto do recurso. Recomendamos que o valor seja selecionado de um vocabulário controlado ou de um esquema formal de classificação.
Nome do elemento: Descrição	
Rótulo:	Descrição
Definição:	Descrição do conteúdo do recurso.
Comentário:	Exemplos de Descrição incluem, mas não se limitam a: um resumo, tabela de conteúdos, referência a uma representação gráfica do conteúdo ou uma descrição em texto livre do conteúdo.
Nome do elemento: Editor	
Rótulo:	Editor
Definição:	Entidade responsável por disponibilizar o recurso.
Comentário:	Exemplos de Editor incluem uma pessoa, uma organização ou um serviço. Normalmente, o nome do Editor deve ser utilizado para indicar a entidade.
Nome do elemento: Colaborador	
Rótulo:	Colaborador
Definição:	Entidade responsável por fazer contribuições ao conteúdo do recurso.
Comentário:	Exemplos de Colaborador incluem uma pessoa, uma organização ou um serviço. Normalmente, o nome do Colaborador deve ser utilizado para indicar a entidade.
Nome do elemento: Data	
Rótulo:	Data
Definição:	Data do evento no ciclo de vida do recurso.
Comentário:	Normalmente, a Data estará associada com a criação e disponibilização do recurso. Recomendamos a utilização da ISO 8601 [W3CDTF] que inclui datas na forma de AAAA-MM-DD (dentre outras).
Nome do elemento: Tipo	
Rótulo:	Tipo do Recurso
Definição:	Natureza ou gênero do conteúdo do recurso.
Comentário:	O Tipo inclui termos que descrevem categorias gerais, funções, gêneros ou níveis de agregação do conteúdo do recurso. Recomendamos que o valor seja selecionado de um vocabulário controlado (por exemplo o *DCMI Type Vocabulary* [DCT1]). Para descrever a manifestação física ou digital do recurso, utilize o elemento FORMATO.
Nome do elemento: Formato	
Rótulo:	Formato
Definição:	Manifestação física ou digital de um recurso.
Comentário:	Normalmente, o Formato pode incluir o tipo de mídia ou dimensões do recurso. O Formato pode ser utilizado para identificar o *software*, *hardware* ou outro equipamento necessário para exibir ou operar o recurso. Exemplos de dimensões incluem tamanho e duração. Recomendamos que o valor seja selecionado de um vocabulário controlado (como, por exemplo, a lista *Internet Media Types* [MIME] que define os formatos de mídia de computador).

Nome do elemento: Identificador	
Rótulo:	Identificador do Recurso
Definição:	Referência não ambígua do recurso num dado contexto.
Comentário:	Recomendamos que o recurso seja identificado através de uma seqüência de caracteres (*string*) ou número de acordo com um sistema formal de identificação. Sistemas formais de identificação incluem, mas não se limitam a: *Uniform Resource Identifier* (URI) (incluindo o *Uniform Resource Locator* (URL)), o *Digital Object Identifier* (DOI) e o *International Standard Book Number* (ISBN)[29].
Nome do elemento: Fonte	
Rótulo:	Fonte
Definição:	Referência ao recurso da qual o recurso é originado.
Comentário:	O recurso pode ser originado de uma Fonte no todo ou em parte. Recomendamos que a referência seja identificada mediante o uso de uma seqüência de caracteres (*string*) ou número de acordo com um sistema formal de identificação.
Nome do elemento: Idioma	
Rótulo:	Idioma
Definição:	Idioma do conteúdo intelectual do recurso.
Comentário:	Recomendamos utilizar o RFC 3066 [RFC3066] que, juntamente com a ISO639 [ISO639], definem *tags* de 2 a 3 idiomas principais com *subtags* opcionais. Exemplos incluem "en" ou "eng" para *English* (Inglês), "akk" para "Akkadian" e "en-GB" para *English* (Inglês) utilizado no Reino Unido.
Nome do elemento: Relação	
Rótulo:	Relação
Definição:	Referência a um recurso relacionado.
Comentário:	Recomendamos que a idetificação de um recurso relacionado seja feita mediante o uso de uma seqüência de caracteres (*string*) ou número de acordo com um sistema formal de identificação.
Nome do elemento: Cobertura	
Rótulo:	Cobertura
Definição:	Extensão ou abrangência do conteúdo do recurso.
Comentário:	Normalmente, a Cobertura incluirá localização especial (nome de um lugar ou coordenadas espaciais), período de tempo ou jurisdição (como uma entidade administrativa). Recomendamos a utilização de um vocabulário controlado para a seleção de uma valor (por exemplo, o *Thesaurus of Geographic Names* [TGN]) e a utilização preferencial, onde apropriado, de nomes de lugares ou períodos de tempo, ao invés de identificadores numéricos, tais como conjuntos de coordenadas ou períodos de tempo.
Nome do elemento: Direitos Autorais	
Rótulo:	Gerenciamento de Direitos Autorais
Definição:	Informação sobre Direitos Autorais do recurso.
Comentário:	Normalmente, o elemento Direitos Autorais conterá uma declaração de direitos para o recurso ou fará referência a um serviço que contenha esta informação. A informação sobre Direitos Autorais inclui *Intellectual Property Rights* (IPR), *Copyright* e vários *Property Rights*. Se o elemento estiver ausente, nenhuma suposição pode ser feita sobre os Direitos Autorais do recurso.

Fonte: *Dublin Core Metadata Element Set, Version 1.1: Reference Description. Disponível em: http://dublincore.org/documents/2003/02/04/dces. Acesso em: 28.07.04.*

[29] Os chamados *Identifiers* (Idenficadores) identificam de forma única uma entidade bibliográfica e, na sua maioria, são designados por uma autoridade responsável por manter a consistência do sistema de identificadores. Podemos citar, como exemplo, as *International Standard Book Numbers* (ISBNs), em que a autoridade responsável é *ISBN Agency* americana.

O esquema não é preso a um único formato e nem a um único conjunto de regras de conteúdo, embora nos comentários acerca dos elementos, como pode ser visto no Quadro 5, em alguns casos um conjunto específico de regras de conteúdo é recomendado. Na *homepage* do DCMI podemos acessar também um guia oficial (*usage guide*) para uso do esquema e que traz mais recomendações. Segundo este guia, o padrão Dublin Core apresenta dois níveis: simples e qualificado (*Qualified*). O primeiro contém 15 elementos, como se vê no Quadro 5, e o segundo, além destes, apresenta um outro elemento denominado Audiência, definido como uma classe de entidade para a qual se destina o recurso como, por exemplo, audiência=professores do ensino fundamental. O Dublin Core Qualificado possui um grupo de qualificadores (*qualifiers*) que tem por objetivo identificar o esquema utilizado para representar um elemento do Dublin Core, assim como refinar seu significado. O elemento Data, por exemplo, apresenta os qualificadores de refinamento: Criado, Válido, Disponível, Modificado, Data de aceitação e Data de submissão, e possui dois esquemas qualificadores: *DCMI period* e o *W3C-DTF*.

Como já analisado, o Dublin Core foi criado para a descrição de recursos em nível básico, de forma que os elementos pudessem ser definidos pelos próprios autores dos documentos, sem que fosse necessário a atuação de um catalogador ou indexador. (WEBER, 2002). A simplicidade para utilização e criação dos elementos do Dublin Core é um dos argumentos utilizados para criticar o padrão MARC que, ao contrário, apresenta complexidade de processamento e de regras. É importante entender que a pretensa complexidade pode também ser característica das muito modernas representações em XML da mesma informação, como atesta Caplan (2003). A despeito das críticas, a autora salienta a importância do Padrão MARC, que tornou possível o compartilhamento bem sucedido de registros catalográficos entre bibliotecas, mais facilmente do que outras áreas de negócios foram capazes de fazer.

O artigo de autoria de Medeiros (1999) cujo título é bem sugestivo, "Making Room for MARC in a Dublin Core World" (Arranjando lugar para o MARC num mundo Dublin Core), a autora relata que uma das discussões controversas era de que o Dublin Core seria utilizado em substituição ao MARC:

> Contudo, para os bibliotecários, o pensamento de abandonar este padrão reconhecido, que tem milhões de registros já investidos, é heresia. Apesar disso, a necessidade de estabelecer e melhorar o acesso aos recursos eletrônicos, combinada como o alto custo da catalogação da *Internet* utilizando o tradicional MARC, fez com que as atenções se voltassem para o Dublin Core (MEDEIROS, 1999, p. 1).

Segundo Medeiros (1999), numa tentativa de integrar as duas visões, a OCLC, sempre pioneira em suas iniciativas, criou o projeto *Cooperative Online Resource Catalog* (CORC)[30], cujo objetivo é usar tanto os registros MARC como os registros Dublin Core para criar uma base de dados de qualidade, composta de recursos da *Internet*.

Reproduzimos, no Quadro 6, as características gerais do Dublin Core e do MARC.

Quadro 6 - Características gerais do Dublin Core e do MARC

DUBLIN CORE	**FORMATO MARC**
15 elementos.	Inúmeros elementos.
Pode ser utilizado por leigos, como também por catalogadores experientes.	Requer treinamento especializado.
Conjunto de elementos comumente compreendidos, o que aumenta a possibilidade de interoperabilidade entre as disciplinas.	Dados conhecidos e compreendidos internacionalmente.
Construído com base em consenso internacional.	Realinhamento do MARC ocorrendo internacionalmente.
Adequado para a descrição de recursos na *Web*.	Melhor adequação para a descrição de recursos impressos e boa adequação para descrição de recursos numa forma física, tangível.
Flexível.	Grande flexibilidade decorrente da integração do formato.
Sem limites para o comprimento do campo.	Mudança e desenvolvimento lentos; ultrapassados pela tecnologia; com freqüência não oferecendo meios adequados para descrever e acessar os recursos na web.
Todos os campos opcionais e repetitivos, quando necessário.	Alguns campos opcionais, outros obrigatórios, apenas alguns são repetitivos.

Fonte: WEBER, Mary Beth. Cataloging Nonprint and Internet Resources: a How-To-Do-It Manual for Librarians. New York: Neal-Schuman Publishers, 2002, p. 356.

Na Figura 7 apresentamos um exemplo de registro Dublin Core, e o mesmo registro em formato MARC.

30 Antes de 30 de Junho de 2002, o CORC era um serviço separado para catalogação e gerenciamento de recursos eletrônicos. As funções do CORC são agora parte da *OCLC Connection*, um serviço que fornece funções de catalogação e acesso ao *WorldCat*.

Figura 7 - Exemplo de registro em Dublin Core e em formato MARC

Registro MARC para o Retrato de Woodcut de John Muir

100	1	McCurdy, Michael
245	10	[Woodcut portrait of John Muir] ‡h[electronic resource]/‡cMichael McCurdy.
245	30	Woodcut of John Muir ‡h[electronic resource]
260		[S.l. : ‡b.n., ‡c19- - ?]
500		World Wide Web resource (viewed on August 3, 2001).
500		Title supplied buy cataloger.
520		Woodcut portrait of Sierra Club founder John Muir featured in the online Exhibit "John Muir : Images and Pictures."
600	10	Muir, John, ‡d1838-1914.
610	20	Sierra Club
650	0	Naturalists ‡zUnited States
856	41	‡chttp://www.sierraclub.org/john_muir_exhibit/pictures/graphics/woodcut_portrait_of_john_muir_by_michael_mccurdy.jpg

Registro Dublin Core para o Retrato de Woodcut de John Muir

Title	Woodcut portrait by Michael McCurdy.
Subject	Woodcut portrait of John Muir by Michael McCurdy
Subject	Muir, John, 1838-1914
Description	Woodcut portrait by artist Michael McCurdy of Sierra Club founder John Muir.
Creator	Wood, Harold.
Publisher	Sierra Club.
Contributor	McCurdy, Michael.
Type	Image
Format	JPEG image (file size = 22 kilobytes)
Identifier	http://www.sierraclub.org/john_muir_exhibit/pictures/graphics/woodcut_portrait_of_john_muir_by_michael_mccurdy.jpg
Source	John Muir Exhibit : Images and Pictures.
Language	English

Fonte: WEBER, Mary Beth. Cataloging Nonprint and Internet Resources: a How-To-Do-It Manual for Librarians. New York: Neal-Schuman Publishers, 2002, p. 358.

6.2 Mapeamento e análise de esquemas de metadados no exterior

Além do Dublin Core e do MARC, há uma infinidade de outros esquemas de metadados utilizados internacionalmente. Uma parte deste universo foi retratado em nossa pesquisa como um dos resultados da análise empírica e também para atender a um dos seus objetivos específicos. O mapeamento considerou 27 esquemas e teve como objetivo retratar a pluralidade de esquemas, considerando a diversidade de comunidades atendidas e áreas e suas diferentes aplicações. Para cada um dos esquemas foram coletadas as seguintes informações: definição/objetivo, instituições responsáveis, comunidades atendidas, *homepage* do esquema e *URLs* para acesso aos elementos relacionados. Estas informações foram coletadas nos *sites* oficiais de cada um dos esquemas, quando identificadas nas *homepages* analisadas. O mapeamento dos esquemas pode ser encontrado no Anexo 1.

O Quadro 7 apresenta a sistematização dos esquemas considerados neste mapeamento, agrupados em **Gerais**, **Especializados por Área** e **Aplicações**.

Quadro 7 - Esquemas de metadados no exterior

CATEGORIA	USO	ESQUEMAS
Metadados gerais	*Geral*	*Dublin Core*
Metadados especializados Por área	*Arte e Arquitetura*	*Categories for the Description of Works of Art (CDWA)*
		Visual Resources Association Core (VRA Core)
	Biologia e Meio Ambiente	*MBII Biological Metadata (NBII)*
		Ecological Metadata Language (EML)
		Darwin Core
	Ciências Sociais	*Data Documentation Initiative (DDI)*
	Educação	*Gateway to Educational Materials (GEM)*
		Learning Object Metadata (LOM)
		Sharable Content Object Reference Model (SCORM)
	Geociências	*Content Standard for Digital Geospatial Metadata (CSDGM)*
	Lingüística e Literatura	*Text Encoding Initiative (TEI)*
Aplicações de metadados	*Arquivos*	*Encoded Archival Description (EAD)*
	Bibliotecas	*Machine Readable Cataloging (MARC)*
	Comércio de Livros	*ONIX International*
	Direitos Autorais	*Interoperability of Data in E-Commerce Systems (INDECS)*
		Open Digital Rights Language (ODRL)
		eXtensible Rights Markup Language (XRML)
	Informação do Governo	*Government Information Locator Service (GILS)*
	Estruturados	*Exchange Format For Electronic Components and Texts (EFFECT)*
		Berkeley Electronic Binding Project (EBIND)
		Making of America II (MOA2)
		MPEG-7 (ISO/IEC 15938)
	Preservação	*Open Archival Information Systems (OAIS)*
	Tecnologia de Áudio e Vídeo	*Technical Metadata for Digital Still Images (TMDSI)*
		Audio Technical Metadata Extension Schema (AUDIOMD)
		Video Technical Metadata Extension Schema (VIDEOMD)

No Quadro 7, os esquemas de Metadados foram divididos, conforme já explicado, em três diferentes categorias: **Metadados Gerais**, que atendem a várias comunidades de diferentes áreas e com diversas aplicações; **Metadados Especializados por Área**, com destaque para as áreas de Arte e Arquitetura, Biologia e Meio Ambiente, Ciências Sociais, Geociências, Lingüística e Literatura; e **Aplicações de Metadados** para Arquivos, Bibliotecas, Comércio de Livros, Direitos Autorais, Informação do Governo, Metadados Estruturados, Preservação e Tecnologia de Áudio e Vídeo.

A sistematização deixa evidente que os esquemas por área referem-se a disciplinas ou campos do conhecimento, enquanto aqueles de aplicação voltam-se a setores, organismos e contextos, conforme distinção de Pinheiro (1999, p. 176)

Como pode ser visto no Quadro 7, há pluralidade de esquemas de metadados, o que nos remete à afirmação de Milstead e Feldman (1999), que se referem a esta característica como "atmosfera caótica de padrões". Na sua maioria, os esquemas coletados em nossa pesquisa podem ser considerados resultados de esforços para integração entre instituições representantes de uma mesma área ou voltadas para uma aplicação. A maioria destas iniciativas ocorre nos Estados Unidos, mas também podem ter abrangência internacional, como é o caso do *Text Encoding Initiative* (TEI).

O Quadro 8 apresenta os criadores/mantenedores dos esquemas de metadados pesquisados.

Quadro 8 - Esquemas de Metadados e seus criadores/mantenedores

Esquemas	*Instituições Responsáveis*
Dublin Core	*Dublin Core Metadata Initiative*
Categories for the Description of Works of Art (CDWA)	*Art Information Task Force (AITF),* projeto *do College Art Association of America e do GETTY Information Institute*
Visual Resources Association Core (VRA Core)	*Visual Resources Association Data Standards Committee*
MBII Biological Metadata (NBII)	*Biological Data Working Group do Federal Geographic Data Committee (FGDC)* e a *Biological Resources Division do United States Geological Survey (USGS)*
Ecological Metadata Language (EML)	*Knowledge Network for Biocomplexity (KNB)*
Darwin Core	*Z39.50 Biology Implementors Group (CBIG),* Projeto *Especies Analyst e o Natural History Museum and Biodiversity Research Center da Kansas University*
Data Documentation Initiative (DDI)	*International Association of Social Science Information Service and Technology (IASSIST), Inter-university Consortium for Political and Social Research (ICPSR), Council of European Social Science Data Services (CESSDA), International Federation of Data Organizations (IFDO), Roper Center for Public Opinion Research*
Gateway to Educational Materials (GEM)	*United States Education Department*
Learning Object Metadata (LOM)	*Institute of Electrical and Electronics Engineers (IEEE), Computer Society, Learning Technology Standards Committe.*
Sharable Content Object Reference Model (SCORM)	*Advanced Distributed Learning Network.*
Content Standard for Digital Geospatial Metadata (CSDGM)	*Federal Geographic Data Committee (FGDC)*
Text Encoding Initiative (TEI)	*Association for Computers and the Humanities, Association for Computational Linguistics* e *Association for Literary and Linguistic Computing*
Encoded Archival Description (EAD)	*Library of Congress* em parceria com a *Society of American Archivists*
Machine Readable Cataloging (MARC)	*Library of Congress*
ONIX International	*Advanced Distributed Learning Network.*
Interoperability of Data in E-Commerce Systems (INDECS)	*Indecs Framework Ltd.*
Open Digital Rights Language (ODRL)	*IPR Systems Pty Ltd.*
eXtensible Rights Markup Language (XRML)	*ContentGuard.*
Government Information Locator Service (GILS)	*United States Government*

Exchance Format For Electronic Components and Texts (EFFECT)	*Elsevier Science*
Berkeley Electronic Binding Project (EBIND)	*University of California* e *Berkeley Library*
Making of America II (MOA2)	*Integrantes da Digital Library Federation: University of California, Berkeley Library, Cornell University, New York Public Library, Pennsylvania State University* e *Standford University*
MPEG-7 (ISO/IEC 15938)	*Motion Picture Experts Group*
Open Archival Information Systems (OAIS)	*Online Computer Library Center (OCLC), Research Libraries Group (RLG)* e *Working Group on Preservation Metadata*
Technical Metadata for Digital Still Images (TMDSI)	*National Information Standards Organization e AIIM International*
Audio Technical Metadata Extension Schema (AUDIOMD)	*Library of Congress*
Video Technical Metadata Extension Schema (VIDEOMD)	*Library of Congress*

Como pode ser visto pelo Quadro 8, os esquemas de metadados são quase sempre oriundos de projetos agenciados por uma ou mais instituições, sendo na sua maioria instituições de pesquisa e universidades.

Em nossa análise sobre os esquemas, foi possível encontrar algumas singularidades, apontadas a seguir.

- O EAD foi o primeiro padrão desenvolvido para a descrição de Guias de arquivos. A comunidade arquivística ressentia-se da falta de um padrão para descrição de coleções de arquivo, de acordo com os princípios da Arquivologia, como o da proveniência e da ordem original dos documentos.
- Há exemplos de esquemas onde temos o envolvimento de várias comunidades voltadas para um mesmo fim, como é o caso dos esquemas para Direitos Autorais e, em especial, o INDECS, pela sua abrangência, do qual participam produtores de filme, gravadoras e editoras de livros e revistas.
- O padrão *Governement Information Locator Service* (GILS) é uma iniciativa do Governo Americano para a descrição de recursos, mas também pode ser utilizado para a descrição de informações não-governamentais e, neste sentido mais amplo, é chamado de *Global Information Locator Service*.
- Os esquemas para a descrição de materiais educacionais pesquisados por nós se diferenciam por descreverem diferentes tipos de materiais. O foco do GEM, por exemplo, é a descrição de planos de aulas, ementas de disciplinas e outros recursos curriculares, enquanto que o LOM e o SCORM descrevem recursos de aprendizagem.

- Há casos em que a partir de um esquema, surge outro. Isto ocorre com o *NBII Biological Metadata Standard* que é, na verdade, um braço do FGDC, denominado perfil (*profile*).
- O DDI é um grupo internacional de produtores de dados em Ciências Sociais, cujo foco é a pesquisa nesta área. O esquema por eles desenvolvido tem o mesmo nome e descreve o conjunto de dados em Ciências Sociais que incluem dados de censo, resultados de pesquisa e estatísticas de saúde, por exemplo.
- O TMDSI , o AUDIOMD e o VIDEOMD são denominados metadados técnicos pois documentam unicamente a criação e as características de arquivos digitais.
- Os esquemas EFFECT, EBIND, MOA2 e MPEG-7 são denominamos metadados estruturados porque descrevem a organização interna de um recurso. No ambiente digital, os recursos digitais são constituídos por diversos arquivos e estes metadadados são necessários para relacionar um arquivo físico a outro, de forma a permitir a estruturação lógica do objeto. Um exemplo é o que ocorre com um livro digitalizado de 100 páginas: cada página é um arquivo imagem com extensão TIFF e os metadados estruturados são utilizados para indicar qual arquivo extensão TIFF é página 1, qual é a 2, e assim sucessivamente.

Pela nossa análise, podemos enfatizar que a preocupação com o desenvolvimento de padrões para o compartilhamento de informações entre estas comunidades remonta à década de 90, período em que a maioria destes esquemas foi desenvolvido, coincidindo com o momento em que o uso da rede realmente se consolidou e se ampliou, no mundo inteiro, inclusive no Brasil. É importante destacar que os esquemas já nascem com foco na interoperabilidade, procurando sempre o consenso nas comunidades em que atuam, tendo como objetivo último compartilhar informações entre eles. No caso do ONIX International, por exemplo, o compartilhamento de informações visa também facilitar a distribuição de produtos agilizando, desta forma, o negócio.

6.3 Mapeamento e análise de esquemas de metadados no Brasil

Além do mapeamento desses esquemas, em âmbito internacional, fizemos uma pesquisa através de questionário (Anexo 2) para verificar a utilização de metadados em serviços brasileiros de informação na *Web* e seus respectivos sistemas de recuperação da informação, seguindo os procedimentos metodológicos descritos no capítulo 1 desta dissertação. A seguir é apresentado o quadro geral da pesquisa, complementado pela tabulação dos itens do questionário, na ordem de sua formulação.

6.3.1 Quadro geral da pesquisa

Para a coleta dos dados foram enviados 35 questionários e 10 questionários foram respondidos, conforme a Tabela 1. O Anexo 3 traz a lista completa das instituições que integraram o universo da pesquisa, contendo apenas os dados cadastrais, tendo sido omitidos os nomes dos respondentes.

Tabela 1 - Quadro geral da pesquisa

Questionários	Número	Percentual
Enviados	35	100%
Recebidos	11	31,4%
Amostra	10	28,6%

A Tabela 1 indica o percentual de respostas recebidas, 31,4%, totalizando 10 questionários, que serviram de amostra para a pesquisa. A amostra é representativa de forma a atender às exigências estatísticas. A partir dos dados foi possível fazer algumas observações e também confirmar tendências por nós já apontadas na pesquisa, no que diz respeito ao conhecimento de metadados, esquemas e sua utilização no Brasil.

Quanto ao perfil dos entrevistados, a maioria é de bibliotecários, exercendo função de coordenação da biblioteca central/sistema de informação de bibliotecas.

Para o universo estudado, coletamos também informações sobre os sistemas de informação/*softwares* de gerenciamento, através da navegação nos *sites* das instituições, conforme apresentados no Quadro 9.

Quadro 9 - Sistemas de informação/*softwares* de gerenciamento das instituições pesquisadas

Instituições	Nome do Sistema de Informação	*Software* de Gerenciamento
PUCSP	LUMEN	ALEPH
PUCMinas	Sistema de Bibliotecas PUC-Minas	PERGAMUM
PUCPR	SIBI-PUCPR	PERGAMUM
UFAM	Sistema de Bibliotecas	PERGAMUM
UFRJ	SIBI/UFRJ	ALEPH
UFRN	SISBI	ALEPH
UFU	SISBI	VTLS
UNB	Sistema de Bibliotecas	PERGAMUM
UNEB	Sistema Integrado de Bibliotecas	*Sotware* da POTIRON
USP	SIBi/USP	ALEPH

Como podemos notar pelo Quadro 9, os sistemas de bibliotecas utilizam em sua maioria, os *softwares* de gerenciamento ALEPH e PERGAMUM, que normalmente contemplam, de forma integrada, as principais funções de uma biblioteca, desde a aquisição até o empréstimo. É digno de nota que o PERGAMUM é um sistema de gerenciamento de bibliotecas nacional, desenvolvido pela Divisão de Processamento de Dados da PUC-PR e que é utilizado, não somente pelas bibliotecas que integram nossa amostra, como também por diversas bibliotecas no país inteiro.

6.3.2 Conhecimento sobre metadados

A Tabela 2 demonstra em que instituições os profissionais de informação conhecem metadados e quais os que forneceram as definições de metadados solicitadas no questionário.

Tabela 2 - Conhecimento e definição de metadados

Instituições	Conhecimento		Definição	
	SIM	NÃO	SIM	NÃO
PUCSP		x		x
PUCMinas	x		x	
PUCPR	x		x	
UFAM	x		x	
UFRJ	x		x	
UFRN	x		x	
UFU	x			x
UNB	x		x	
UNEB	x		x	
USP	x		x	

Podemos concluir que a maioria dos estrevistados tem conhecimento de metadados e apresentou definições extraídas da literatura para representar suas idéias, com exceção da PUCSP.

As definições de metadados apontadas pelos responsáveis por sistemas de informação, encontram-se transcritas a seguir:

PUCMinas: "Segundo SHAEFER (1998), os metadados são importantes para a identificação, organização e recuperação da informação digital. Sua finalidade é facilitar, globalmente, a localização e recuperação das informações eletrônicas, para os usuários. Neste sentido, utiliza-se procedimentos técnicos de indexação e classificação dos conteúdos informacionais, possibilitando a integração de fontes diversificadas e heterogêneas de informação".

PUCPR: "Descrições de dados armazenados em bancos de dados ou como é comumente definido "dados sobre dados a partir de um dicionário digital de dados". Segundo Sumpter, "Metadado é a informação sobre os dados que permite o acesso e gerenciamento deste dado de maneira eficiente e inteligente".

UFAM: "Metadados são informação que resume, enriquece ou complementa os objetos ou serviços referenciados, produzindo assim um potencial incremento de informação. Dados descritos em padrões internacionalmente aceitos."

UFRJ: "São os dados dos dados, ou seja; é a documentação (eficiente) dos sistemas e bancos de dados que descreve o uso dos recursos eletrônicos, de maneira bibliográfica".

UFRN: "Conjunto de elementos padronizados que possibilita representar as informações eletrônicas e a descrição de recursos eletrônicos de maneira bibliográfica".

UNB: "Meu conhecimento sobre metadados é superficial. Normalmente são definidos como dados sobre dados para incrementar a informação. São elementos retirados de um documento, por exemplo, que descrevem e melhoram a informação sobre este documento. Permite acessar facilmente a informação".

UNEB: "Conjunto de dados estruturados que identificam os dados de um determinado documento e que podem fornecer informação sobre o modo de descrição, administração, requisitos legais de utilização, funcionalidade, técnica, uso e preservação".

USP: "As transformações observadas, em âmbito internacional, na área de biblioteconomia como em outras, decorrem das mudanças freqüentes do cenário sócio-econômico e, em grande parte, do desenvolvimento tecnológico verificado nas últimas décadas. As tendências apontam para maior racionalidade nas ações, com cooperação para compartilhamento de recursos e esforços, na própria instituição ou entre instituições congêneres, com uso de tecnologia de informação. Nesse contexto, a *Internet* transformou-se em importante meio de geração/edição e disseminação de recursos de informação, tais como bases de dados, *websites* etc. Essa prática exige padrões de comunicação e de tratamento de dados, como novos modelos de estruturas dos mesmos: conjunto de elementos – metadados – para descrição, armazenagem e localização do objeto digital com recursos de tecnologia, ampliando a disseminação e o acesso à informação, via redes, no local em que estiver, independente da posse do documento físico. Registre-se, ainda, as facilidades de migração de dados referentes aos recursos de informação, que a adoção de formatos padronizados proporciona".

Embora na maioria das vezes os profissionais de informação tenham indicado definições extraídas da literatura, conforme solicitado, a seleção e a escolha de autores e definições pode demonstrar quais as características, funções ou aspectos são mais facilmente identificados e reconhecidos por estes profissionais.

Para analisarmos estes elementos foi elaborada uma outra tabela (Tabela 3), sistematizando aspectos das definições e fazendo a confluência entre eles, apresentando-os por freqüência, em ordem decrescente, de forma a mostrar primeiro os termos mais citados, seguidos dos menos citados. É importante destacar que a PUCSP e a UFU não constam na Tabela 3, pois os profissionais respondentes, de ambas, não forneceram definições sobre o tema.

Tabela 3 - Confluência de aspectos sobre metadados extraídos das definições

Aspectos	**Instituições**								**Total**
	PUCMinas	PUCPR	UFAM	UFRJ	UFRN	UNB	UNEB	USP	
Dados		x	x	x		x	x	x	6
Descrições		x		x	x	x	x	x	6
Informação	x	x	x			x	x	x	6
Documento				x		x	x	x	4
Acesso		x				x		x	3
Padrões			x		x			x	3
Bancos de dados		x		x					2
Conjunto de elementos					x			x	2
Eficiência		x		x					2
Identificação	x						x		2
Incremento			x			x			2
Informações eletrônicas	x				x				2
Localização	x							x	2
Maneira bibliográfica				x	x				2
Objetos			x					x	2
Recursos eletrônicos				x	x				2
Uso				x			x		2
Administração							x		1
Armazenamento								x	1
Classificação	x								1
Complementação			x						1
Comunicação								x	1
Conteúdos informacionais	x								1
Dados estruturados							x		1
Dicionário digital de dados		x							1
Disseminação								x	1

Elementos						x			1
Enriquecimento			x						1
Facilitação								x	1
Fontes	x								1
Formatos								x	1
Funcionalidade							x		1
Gerenciamento		x							1
Indexação	x								1
Informação digital	x								1
Integração	x								1
Inteligência		x							1
Melhoramento						x			1
Migração de dados								x	1
Modelos de estrutura								x	1
Organização	x								1
Preservação							x		1
Procedimentos técnicos	x								1
Recuperação	x								1
Recursos de informação								x	1
Redes								x	1
Representação					x				1
Requisitos legais							x		1
Resumo			x						1
Serviços			x						1
Sistemas				x					1
Técnica							x		1
Tecnologia								x	1
Tratamento								x	1
Usuários	x								1

De acordo com a Tabela 3, os termos mais citados se relacionam às funções/aplicações dos metadados e também às entidades relacionadas. Os termos referentes às funções/aplicações dos metadados e que refletem a importância de seu uso são: descrição, acesso, padrões, eficiência, identificação, incremento, localização, maneira bibliográfica e uso. Dentre estes termos, o mais citado é "descrição" (freqüência=6), que reforça a idéia de que metadados estão fortemente relacionados à descrição, para os respondentes.

Os seguintes termos estão relacionados às entidades para as quais os metadados são aplicados: dados, informação, documento, banco de dados, conjunto de elementos, informações eletrônicas, objetos e recursos eletrônicos. Dentre estes termos, os mais citados são "dados" e "informação" (freqüência=6), o que reforça a definição de metadados como dados sobre dados e informação, objeto de estudo da Ciência da Informação.

O conjunto de todos os termos da Tabela 3 comprova a riqueza de aspectos dos metadados referentes à sua definição, aplicação e atributos, e demonstra a complexidade da questão.

6.3.3 Conhecimento sobre esquemas de metadados

O Quadro 10 apresenta os esquemas de metadados conhecidos pelos entrevistados, sendo os mais conhecidos o MARC e o Dublin Core.

Quadro 10 - Esquemas de metadados conhecidos

Instituições	Esquemas de Metadados
PUCSP	-
PUCMinas	Dublin Core e MARC
PUCPR	Dublin Core e MARC
UFAM	IEEE/LOM e MARC
UFRJ	Dublin Core e MARC
UFRN	MARC
UFU	MARC
UNB	Dublin Core e MARC
UNEB	Dublin Core e MARC
USP	Dublin Core e MARC e GILS

Pelo universo estudado, qual seja, o das bibliotecas universitárias, é natural que o MARC, formato mais utilizado em bibliotecas, fosse o mais citado de todos, com 09 referências. Mas é importante entender que a utilização do termo metadados é recente e que por isso pode-se considerar também significativo o conhecimento demonstrado a respeito do padrão Dublin Core, segundo mais citado, com 06 ocorrências. É compreensível que este padrão, dentre tantos outros, seja internacionalmente reconhecido por diversas comunidades, não se limitando ao mundo das bibliotecas, por seu caráter de ampla aplicabilidade e abrangência. Aparecem também o padrão IEEE/LOM, que atende à descrição de recursos educacionais e o GILS, voltado para a descrição de recursos de informação do Governo Americano. Maiores detalhes sobre estes padrões podem ser encontrados no mapeamento dos esquemas no exterior, no Anexo 1.

6.3.4 Utilização de metadados e especificação dos esquemas

A Tabela 4 mostra a utilização ou não de esquemas de metadados pelos sistemas de bibliotecas e quais são eles.

Tabela 4 - Uso de metadados e esquemas utilizados

Instituições	Uso		Esquema
	SIM	NÃO	
PUCSP	-	-	-
PUCMinas	x		MARC
PUCPR	x		MARC
UFAM	x		MARC
UFRJ		x	-
UFRN	x		Esquema Local
UFU	x		MARC
UNB		x	-
UNEB	x		MARC
USP	x		MARC e Dublin Core

A Tabela 4 mostra que o MARC é o padrão mais utilizado pelas 10 bibliotecas entrevistadas, seguido do Dublin Core e de um esquema local. A partir destes dados, confirma-se o fato de que o padrão MARC continua sendo preponderante, pelo menos no Brasil, na comunidade das bibliotecas e que esta tendência parece ser muito forte, apesar das desvantagens apontadas por alguns autores e comentadas nesta dissertação.

A tradição do uso do MARC no Brasil remonta ao Projeto CALCO (Catalogação Legível por Computador), criado com o objetivo de intercambiar informações catalogadas entre bibliotecas, marco da catalogação cooperativa no país e baseado inteiramente no formato MARC II. Importante notar que o CALCO foi resultado dos estudos da Professora Alice Príncipe Barbosa, para sua dissertação de mestrado no IBBD (Instituto Brasileiro de Bibliografia e Documentação), atual IBICT (Instituto Brasileiro de Informação em Ciência e Tecnologia). Em 1975 foi decidido que o formato CALCO seria adotado em nível nacional para o processamento de dados bibliográficos referentes à produção bibliográfica brasileira, resolução esta tomada pelo IBBD em reunião de especialistas para a implementação dos Sistemas Nacionais de Informação (NATIS), um projeto da UNESCO (BARBOSA, 1978).

Uma das utilizações mais conhecidas do formato CALCO foi a Rede BIBLIODATA/CALCO da Fundação Getúlio Vargas (FGV), que passou a chamar-se apenas rede BIBLIODATA, quando a FGV decidiu fazer a conversão de seus registros bibliográficos para o USMARC, no período de 1994 a 1996, pois o CALCO, embora tenha sido baseado no padrão MARC, na época em que foi criado, foi ficando muito defasado em relação a este ao longo do tempo.

A tendência de continuação do uso do formato MARC pelas bibliotecas é apontada por Medeiros (1999), quando a autora diz que em muitas organizações, os catálogos *online*, por exemplo, continuarão a ser representados no formato MARC. Mas, por outro lado, em nossa

pesquisa, temos o caso da USP, em que o MARC é utilizado para seu Banco Bibliográfico, o DEDALUS, enquanto que "para outros recursos de informação eletrônica, disseminadas em *website*, foi definido um conjunto de elementos para o SIBI/USP, com base no modelo de estrutura de dados Dublin Core" (respondente da USP). Muito provavelmente isto ocorre porque, como aponta Medeiros (1999), a catalogação de alguns recursos eletrônicos em formato MARC, que é um formato de representação robusto, pode não ser mais justificada, sendo substituído pelo Dublin Core. Em seu artigo, a autora conclui que "o MARC e o Dublin Core combinados são maiores que a soma de suas partes. Através de uma relação de complementariedade, estes dois padrões descritivos podem possibilitar o acesso tão necessitado ao que de melhor a *Net* tem a oferecer" (MEDEIROS, 1999, p. 3).

7 CONCLUSÃO

Esta dissertação foi iniciada pela descrição das técnicas de indexação, classificação e catalogação por serem os pilares de sustentação ao sistema de recuperação da informação e porque, como atividades de representação da informação, impactam diretamente na sua capacidade de recuperar informações de um sistema. Nesse sentido, a recuperação da informação deve atender às necessidades e demandas dos usuários e ser capaz de localizar, recuperar ou permitir o acesso à informação em grandes conjuntos de documentos, objetos e informações. Ao mesmo tempo, considerando as inter-relações existentes entre essas técnicas - catalogação, classificação e indexação - e sua importância para o sistema de recuperação da informação, o conhecimento sobre os seus primórdios e desenvolvimento evoluiu até os metadados.

Os avanços do sistema de recuperação da informação, por sua vez, foram estudados no âmbito da Ciência da Informação, entendendo que este sistema pode ser considerado o coração da própria área, que lhe dá origem, como raiz, a partir da preocupação em solucionar o problema de recuperar informação em grandes conjuntos de documentos e da disponibilidade de novas tecnologias.

Esta explosão informacional não é um fenômeno atual e tem sido abordada por diferentes autores em épocas distintas e anteriores, tendo sido destacados Paul Otlet e Vannevar Bush e suas idéias revolucionárias.

Nesta dissertação foi enfocado o quanto a evolução dos sistemas de recuperação da informação depende muito dos avanços obtidos nas técnicas e métodos empregados com este objetivo e das tecnologias de informação, quando surgem os sistemas de recuperação *online*. Neste novo ambiente, os critérios utilizados de avaliação do sistema de recuperação da informação e os principais instrumentos de recuperação da informação e suas peculiaridades foram pontos destacados, pela sua relevância.

O estudo do sistema de recuperação da informação e de seus fundamentos teóricos são a base para o entendimento das questões relacionadas à recuperação da informação na *Web*., pois o fenômeno crescente de documentos acontece tanto ou mais no ambiente virtual e há necessidade premente em atender às necessidades de informação de seus usuários, que se sentem frustados ao navegar em suas páginas.

E neste sentido foi analisada a recuperação da informação na *Web*, o que levou à constatação de que os problemas de recuperação da informação neste ambiente não são novos e

que os profissionais das áreas de Biblioteconomia, Museologia, Arquivologia e cientistas da informação têm lutado para solucioná-los, por décadas. Tendo em mente o estado caótico em que se encontra a *Web*, estas questões são ainda mais significativas para uma nova e mais ampla comunidade de usuários.

É tão importante que os profissionais de informação destas áreas assumam estes novos desafios quanto os novos atores não ignorem o legado de conhecimento destes profissionais e o saber construído ao longo do tempo. Estes profissionais que hoje atuam em redes eletrônicas de comunicação e de informação não podem desconhecer o fato de que muitas das questões da recuperação da informação na rede já foram abordadas anteriormente, num ambiente ainda não automatizado e antes da virtualidade e do ciberespaço.

Entendemos que a *Web* é um ambiente onde estas práticas tradicionais podem e devem ser renovadas e (re)utilizadas, como também novos desenvolvimentos e inovações devem ser buscados, pois a rede, por ser multimídia e ter características múltiplas e distintas, apresenta novas perspectivas e uma série de questões até então inexploradas.

É neste contexto, em que novos desafios para o tratamento da informação devem inspirar novas soluções, mas sem relegar as bases sólidas sobre os quais se desenvolveram as práticas de representação e de tratamento ou processamento da informação, que devemos entender o papel desempenhado pelos metadados nos sistemas de recuperação da informação atuais.

Assim, foi abordado como, no ciberespaço, técnicas e metodologias "tradicionais" ou "convencionais" de bibliotecas, tais como as mencionadas catalogação, classificação e indexação estão sendo utilizadas, tendo sido feita uma analogia entre metadados e as técnicas de representação do conteúdo dos documentos. Entendemos que estas técnicas, tão importantes para os sistemas de recuperação da informação, estão sendo utilizadas num novo ambiente, mas com o mesmo objetivo.

Nesta pesquisa foram levantados as diversas definições e conceitos de metadados, seus tipos, características e funções. Concluímos que os metadados são compreendidos de formas diferentes, dependendo da comunidade e do contexto em que são utilizados. De todas as definições, destacamos aquelas nas quais foram consideradas as suas várias aplicações. Além disso, ao analisar o que constitui um esquema de metadados, foram também estudadas suas regras de conteúdo, que normalmente prescrevem o uso de instrumentos como o tesauro ou esquema de classificação, ferramentas tão relevantes para os sistemas de recuperação da informação tradicionais, agora ainda mais importantes para a recuperação da informação no ambiente virtual. Destacamos o papel das sintaxes de transporte de metadados e ressaltamos a

interoperabilidade como um fator de importância primordial para que os metadados possam realmente exercer, ao máximo, suas funções para a recuperação da informação.

Nos resultados de nossa pesquisa empírica apresentamos aspectos do surgimento do Dublin Core e seus efeitos no mundo MARC. A partir do mapeamento de 27 esquemas de metadados no exterior, foi possível confirmar a sua pluralidade pois são muitas comunidades a serem atendidas e estes esquemas atestam os esforços de integração entre as instituições representantes de uma mesma área ou voltadas para determinada aplicação. Concluímos também que a preocupação com o desenvolvimento de padrões é uma constante, principalmente a partir da década de 90, quando a rede se consolidou e ampliou, no mundo inteiro, inclusive no Brasil.

Em contraponto, analisamos a utilização de metadados em sistemas de informação no Brasil e concluímos que:

- há conhecimento sobre metadados, entre profissionais de informação de sistemas acadêmicos;
- metadados estão fortemente relacionados à descrição, predominando o entendimento de metadados referindo-se principalmente à descrição de recursos;
- o padrão MARC é o mais conhecido, sendo reconhecido como metadados pelos profissionais da informação abordados, embora o termo seja recente; na verdade, o padrão MARC é de conhecimento comum no mundo da catalogação "tradicional" das bibliotecas. O conhecimento acerca do Dublin Core, mesmo tendo sido citado em segundo lugar, é também bastante significativo pois este padrão já surgiu enquanto esquema de metadados.
- além de ser mais conhecido, o MARC também é o mais utilizado pelas bibliotecas integrantes da amostra de nossa pesquisa e esta tendência parece ser muito forte; e
- outra tendência verificada foi da utilização dos dois padrões de forma complementar, de acordo com a necessidade; a utilização do formato MARC pode não se justificar em muitos casos, pois é um formato de representação "robusto", sendo substituído pelo Dublin Core.

De um modo geral, nesta dissertação podemos corroborar os argumentos de Gilliland-Swetland (1998) sobre metadados, não somente ao justificar os custos e esforços envolvidos, uma vez que, embora importantes, são um constructo complexo que pode ser caro de criar e manter, mas também em relação à outras constatações mais amplas, principalmente quanto a coleções de museus e arquivos. As idéias de Gilliland-Swetland (1998) foram manifestadas em relação aos seguintes aspectos de metadados:

- *Acessibilidade crescente*: a recuperação é mais eficaz quando utilizados metadados ricos e consistentes, melhorando e possibilitando a busca em várias coleções ou criando coleções virtuais de materiais distribuídas em diversos repositórios.
- *Incorporação de elementos de dados comuns*: os elementos do Encoded Archival Description (EAD), do Text Encoding Initiative (TEI) e do Dublin Core estão sendo incorporados por sistemas de informação digital e por padrões de metadados emergentes, desenvolvidos por diferentes tipos de comunidades profissionais que "estão tornando mais fácil para o usuário negociar entre substitutos descritivos dos objetos de informação e versões digitais dos próprios objetos e buscar tanto no nível do item quanto no nível da coleção dentro e através dos sistemas de informação" (p. 6).
- *Retenção de contexto*: repositórios de museus, arquivos e bibliotecas não retêm simplesmente objetos pois mantêm coleções de objetos que possuem inter-relações complexas entre si, além de associações com pessoas, lugares, movimentos e eventos, isto é, o seu contexto. Assim, não é difícil, no mundo digital, um objeto único de uma coleção depois de digitalizado ser separado, tanto de sua própria informação de catalogação quanto de sua relação com outros objetos na mesma coleção. Neste sentido "os metadados desempenham um papel crítico em documentar e manter estas relações, assim como indicar a autenticidade, integridade estrutural e amplitude dos objetos de informação. Por exemplo, documentar o conteúdo, contexto e estrutura de um registro de arquivo ajuda a distinguir se aquele registro de informação descontextualizada corresponde aos metadados na forma de um Guia de Arquivo" (p. 6).
- *Expansão no uso*: os metadados possibilitam a reprodução digital de documentos e objetos únicos de coleções de arquivos e museus, tornando mais fácil a sua disseminação e acesso universal, por usuários que desta forma podem conhecer uma obra de arte, o que de outra maneira, por dificuldades econômicas, distâncias geográficas e outras barreiras, não seria possível. São novas comunidades de usuários que apresentam necessidades e demandas de informação que diferem muito daquelas de comunidades tradicionais e de especialistas e pesquisadores para os quais foram planejados e implantados os sistemas de informação. "Os metadados podem documentar as mudanças quanto ao uso de sistemas e seu conteúdo e esta informação pode ser um *feedback* importante nas decisões que envolvam o desenvolvimento de sistemas. Metadados bem estruturados podem facilitar um quase infinito número de caminhos para buscar informação, apresentar resultados, e até mesmo manipular objetos de informação sem comprometer sua integridade" (p. 7).

- *Multiplas versões:* "A existência de informação e de objetos culturais em forma digital tem intensificado o interesse na habilidade em criar múltiplas e várias versões daqueles objetos. Este processo pode ser tão simples como criar uma cópia de alta resolução para preservação ou pesquisa quanto uma imagem de baixa resolução que possa ser rapidamente transferida pela rede como referência rápida., ou até mesmo resultar na criação de formas derivadas ou variadas a serem utilizadas em publicações, exibições, ou salas de aula, por exemplo. Em quaisquer dos casos, deve haver metadados que façam a ligação ou que vinculem as múltiplas versões e captem o que é igual e o que é diferente em cada versão. Os metadados devem também ser capazes de distinguir o que é diferente entre versões digitalizadas e a cópia original ou objeto que lhe deu origem" (p. 7).
- *Questões legais*: os metadados podem rastrear as "muitas camadas" relativas a direitos autorais e de reprodução, aspectos existentes nas múltiplas versões de objetos de informação; além disso, os metadados são também capazes de documentar requisitos legais e de propriedade como por exemplo, privacidade e propriedade.
- *Preservação*: os metadados podem ser utilizados para garantir a sobrevivência dos objetos de informação digitais de hoje, em relação às sucessivas gerações de *hardware* e *software* ou à conversão para sistemas completamente novos. Portanto, metadados são necessários a fim de que possam "existir independentemente do sistema que esteja sendo utilizado para armazená-los e recuperá-los... Para que os objetos de informação continuem acessíveis e inteligíveis no decorrer do tempo, será essencial preservar e migrar estes metadados" (p. 8).

Nesta dissertação, uma das polêmicas envolvendo metadados que aparece no seu decorrer é sobre a necessidade ou não da intervenção humana na recuperação da informação. Trata-se do uso de mecanismos de buscas ou da representação de informação, documentos e objetos, na Web, pelo uso de normas de catalogação, classificação e indexação e seus respectivos instrumentos - cabeçalhos de assunto, vocabulários controlados e tesauros.

Os resultados desta dissertação confirmam a necessidade de normas e padrões para a recuperação de informação especializada, por exemplo, informação científica e tecnológica, direcionada a uma determinada comunidade de especialistas, pesquisadores e professores, como é o caso das bibliotecas universitárias abordadas nesta pesquisa.

Assim como Gilliland-Swetland (1998) finalizamos adotando como metáfora a Pedra da Rosetta, da escrita hieroglífica do antigo Egito, decifrada por Champollion, inclusive por análise comparativa. Os hieróglifos não eram utilizados na escrita comum e sim em monumentos e inscrições, tal como os metadados, não são na vida cotidiana e sim uma linguagem, um código

apropriado para "inscrições" de documentos, objetos e informações, pelos profissionais de informação. Os metadados, particularmente na *WWW*, estão em sua infância e certamente vão continuar a evoluir, de tal forma que embora de uso profissional, decifrá-los não exija um grande esforço dos usuários e seja tarefa mais simples. Esta dissertação nos levou a pensar que os profissionais da informação podem e devem utilizar metadados nas suas "inscrições" visando à precisão e consistência no sistema de recuperação da informação e permitindo a universalização de acesso.

O uso contínuo e consistente de esquemas de metadados pode transformar a caótica massa de informação, disponível na rede, numa biblioteca digital ou virtual. E como Champollion na Pedra de Rosetta, profissionais de informação podem fornecer a chave para " decifrar", num mapa delineado para navegação precisa, informações, documentos e objetos dispersos no ciberespaço.

8 REFERÊNCIAS BIBLIOGRÁFICAS

BARBOSA, Alice Príncipe. Novos Rumos da Catalogação. Rio de Janeiro: BNG/Brasilart, 1978.

BELLCORE, Michael Lesk. The Seven Ages of Information Retrieval. Disponível em: http://www.ifla.org/VI/5/op/udtop5/udtop5.htm. Acesso em: 26/07/2004.

BORKO, H. Information Science: what is it ? American Documentation, 19 (1): 3: 5, Jan. 1968.

CAMPOS, Maria Luiza de Almeida.. As Cinco Leis da Biblioteconomia e o Exercício Profissional. Disponível em: <http://www.conexaorio.com/biti/mluiza/index.htm>. Acesso em: 30/03/2004.

CAMPOS, Maria Luiza de Almeida. Linguagem Documentária: Teorias que fundamentam sua elaboração. Niterói; RJ: EdUFF, 2001.

CAPLAN, Priscilla. Metadata Fundamentals for All Librarians. Chicago: American Library Association, 2003.

CARVALHO, Maria Carmem Romcy de. Compartilhamento de Recursos e Acesso à Informação no Brasil: um Estudo das Áreas de Química e Engenharia Química. Brasília: UNB, 1999. Tese (Doutorado em Ciência da Informação).

CASTELLS, Manuel. A Revolução da Tecnologia da Informação. In: Sociedade em Rede (A Era da Informação, vol. I). São Paulo: Paz e Terra, 1999.

CENDóN, Beatriz Valadares. Ferramentas de Busca na *Web*. Ciência da Informação, Brasília, v. 30, n. 1, p. 39-49, jan./abr. 2001.

CHOWDHURY, G. G. The Internet and Information Retrieval Research: a brief review. Journal of Documentation, v. 55, n. 2, p. 209-225, Mar. 1999.

DESIRE. Projeto RE 1004 (RE). The Role of Classification Schemes in Internet Resource Description and Discovery, 19.02.1997.

FROELICH, Thomas J. Caveat Web Surfer ! Responsabilidade Social e Recursos da Internet. Revista Transinformação, São Paulo, v. 10, n. 2, maio/agosto, 1998. Disponível também em <http://www.puccamp.br/~biblio/transinformacao/old/vol10n2/pag15.html>. Acesso em 13/09/2002.

GILL, Tony. Metadata and the World Wide Web. In: Introduction to Metadata: Pathways to Digital Information. California, 1998, p. 9-18.

GILLILAND-SWETLAND, Anne J. Defining Metadata. In: Introduction to Metadata: Pathways to Digital Information. California, 1998, p. 1-8.

GOMES, Hagar Espanha. Bases de Dados Bibliográficos: descrição e representação. Parte 1. Descrição e representação bibliográfica. In: Programa de Treinamento: Aplicação da Tecnologia no Desenvolvimento da Bibliotecas. Rio de Janeiro, Fundação Getúlio Vargas – FGV, 1997. p. 5-8.

GOMES, Hagar Espanha. Uma Profissão de Futuro. Disponível em: <http://www.fgv.br/dg/diti/bib/geral/htm/hpbb12.htm>. Acesso em: 27/10/2000.

GOMES, Sandra Lúcia Rebel. Bibliotecas Virtuais: Informação e Comunicação para a Pesquisa Científica. Orientadora: Lena Vania Ribeiro Pinheiro. Rio de Janeiro: IBICT-UFRJ-ECO, 2002. Tese. (Doutorado em Ciência da Informação).

HARTER, Stephen P. Online Information Retrieval: concepts, principles and techniques. London: Academic Press, 1986.

HUDGINS, Jean, AGNEW, Grace, BROWN, Elizabeth. Getting Mileage out of Metadata: applications for the Library. Chicago: American Library Association, 1998.

KRAEMER, Ligia Leindorf Bartz. Metadados: estudo de sua aplicação no tratamento de recursos virtuais e análise de um projeto do Programa Prossiga do IBICT. Orientadora: Graça Maria Simões Luz. Curitiba: CEFET-PR, 2001. Diss. (Mestrado em Tecnologia).

LANCASTER, F. W. Indexação e Resumos: teoria e prática. Brasília: Briquet de Lemos/Livros, 1993.

LANCASTER, F. W. Information Retrieval Systems: characteristics, testing and evaluation. 2. ed. New York: Wiley-Insterscience, 1979.

LASTRES, Helena M. M. e FERRAZ, João Carlos. Economia da informação, do conhecimento e do aprendizado. In: Informação e Globalização na Era do Conhecimento. Rio de Janeiro: Campus, 1999, p. 27-57.

LÉVY, Pierre. O que é o virtual? Trad. de Paulo Neves. São Paulo: Ed. 34, 1996.

LYMAN, Peter, VARIAN, Hal R. How much information ? Disponível em <http://www.sims.berkeley.edu/projects/how-much-info-2003/internet.htm>. Acesso em: 09/08/2004.

MEDEIROS, Norm. Making room for MARC in a Dublin Core World. Online, November 1999. Disponível em: http://onlinemc.com/onlinemag/OL1999/medeiros11.html. Acesso em: 10.10.03.

MILSTEAD, Jessica, FELDMAN, Susan. Metadata: Cataloguing by any other name. ONLINE, January, 1999. Disponível em: <http://www.infotoday.com/online/OL1999/milstead1.html>. Acesso em: 01/10/2001.

NOVELLINO, Maria Salet Ferreira. A transferência da informação através dos seus contextos de produção e uso: linguagens de transferência da informação. 2000. 167 p. Tese (Doutorado em Ciência da Informação) – Instituto Brasileiro em Informação e Tecnologia, Rio de Janeiro, 2000.

PALMER, Roger C. Online Reference and Information Retrieval. Littleton; Colorado: Libraries Unlimited, 1987.

PEREIRA, Vania Lúcia da Cunha. Sistemas de redução da informação: uma (IR)Recuperação Metodologicamente Configurada. Orientadora: Gilda Maria Braga. Rio de Janeiro: IBICT-UFRJ-ECO, 1994. Diss. (Mestrado em Ciência da Informação).

PEREZ, Antonio Hernández. La búsqueda y recuperación de información em internet. In: La Sociedade de la Información: Política, Tecnologia e Indústria de Contenidos. Madrid: Editorial Centro de Estudos Ramón Areces, 2000.

PIEDADE, Maria Antonieta Requião. Introdução à teoria da classificação. 2. ed. rev. e aum. Rio de Janeiro: Interciência, 1983.

PINHEIRO, Lena Vania Ribeiro. O desafio da formação profissional: da biblioteca às bibliotecas digitais e virtuais. In: INTEGRAR - Congresso Internacional de Arquivos, Bibliotecas, Centros de Documentação e Museus, 1. Textos. São Paulo: FEBAB, 2002, p. 387-404.

PINHEIRO, Lena Vania Ribeiro, LOUREIRO, José Mauro Mattheus. Traçados e limites da Ciência da Informação. Ciência da Informação, Brasília, v. 24, n.1, p. 42-53, jan./abril 1995.

PINHEIRO, Lena Vania Ribeiro. Campos Interdisciplinares da Ciência da Informação: fronteiras remotas e recentes. In: Pinheiro, Lena Vania Ribeiro, org. Ciência da Informação, Ciências sociais e Interdisciplinaridade. Brasília, Rio de Janeiro, IBICT/DEP, 1999, p.155-182.

RIEUSSET-LEMARIÉ, Isabelle. P. Otlet´s Mundaneum and the International Perspective in the History of Documentation and Information Science. In: HAHN, T. B. & Buckland, M. Historical Studies in Information Science. Medford, NJ: ASIS, p. 34 - 42, 1998.

ROBREDO, Jaime, CUNHA, Murilo B. da, colab. Documentação de hoje e de amanhã: uma abordagem informatizada da biblioteconomia e dos sistemas de informação. 2. ed. rev. e ampl. Brasília, Edição de Autor, 1986.

ROSETTO, Márcia. Uso do Protocolo Z39.50 para recuperação de informação em redes eletrônicas. Ci. Inf. vol.26 n.2 Brasília May/Aug. 1997.

ROSETTO, Marcia, NOGUEIRA, Adriana Hypolito. Aplicação de Elementos Metadados Dublin Core para descrição de dados bibliográficos on-line da Biblioteca Digital de Teses da USP. Disponível em http://acd.ufrj.br/sibi/snbu/snbu2002/oralpdf/82.a.pdf. Acesso em: 03/02/2003.

SARACEVIC, Tefko. Interdisciplinary Nature of Information Science. Ciência da Informação, Brasília, v. 24, n. 1, p. 36-41, jan./abril 1995.

SARACEVIC, Tefko. Information Science: origin, evolution and relations. In: VAKKARI, Pertti, CRONIN, Blaise, ed. Conceptions of Library and Information Science: historical, empirical and theoretical perspectives. Proceedings of the International Conference held for the celebration of the 20th Anniversary of the Department of Information Studies. University of Tampere, Finland, 26-28, August 1991. London, Los Angeles: Taylor Graham, 1992, p. 5 –27.

SAYÃO, Luiz Fernando. Bases de dados e suas qualidades. In: Informação e Informática. Salvador: EDUFBA, 2000, p. 143-180.

SCHELLENBERG, Theodore Roosevelt. Documentos públicos e privados: arranjo e descrição. Rio de Janeiro: Editora da Fundação Getúlio Vargas, 1980.

SCHWARTZ, Candy. *Web* Search Engines. In: Journal of American Society for Information Science, 49 (11): 973-882, 1998.

SOUZA, Marcia Izabel Fugiwasa, VENDRÚSCULO, Laurimar Gonçalves, MELO, Geane Cristina. Metadados para a descrição de recursos de informação eletrônica: utilização do padrão Dublin Core. Ciência da Informação, v. 29, n. 1, Brasília. Jan./Abril 2000.

SOUZA, Renato Rocha de, ALVARENGA, Lídia. A *Web* Semântica e suas contribuições para a Ciência da Informação. Ciência da Informação, Brasília, v. 33, n. 1, 2004.

SOUZA, Rosali Fernandez de. A Classificação como Interface da Internet. DataGramaZero – Revista de Ciência da Informação – v. 2, n. 2, abr/00.

SOUZA, Terezinha Batista de, CATARINO, Maria Elisabete, SANTOS, Paulo Cesar. Metadados: catalogando dados na Internet. Revista Transinformação, São Paulo, v. 9, n. 2, maio/agosto, 1997.

WEBER, Mary Beth. Cataloging Nonprint and Internet Resources: a How-To-Do-It Manual for Librarians. New York: Neal-Schuman Publishers, 2002.

WEIBEL, S. , KOCH, Traugott. The Dublin Core Metadata Initiative: mission, current activities and future directions. D-Lib Magazine, v. 6, n. 12, Dezembro 2000. Disponível em: http://www.dlib.orgt/dlib/december00/weibel/12weibel.html. Acesso em: 19/07/02.

WEIBEL, S., KUNZE, J., LAGOZE, C., WOLF, M. Dublin Core Metadata for Resource Discovery. IETF #2413. The Internet Society, September, 1998. Disponível em: http://www.ietf.org/rfc/rfc2413.txt. Acesso em: 28.07.04.

WOODWARD. Cataloging and Classifying Information Resources on the Internet. In: Annual Review of Information Science and Technology (ARIST), v. 31, 1996, p. 189- 220.

L I S T A D E S I T E S

Art, Design, Architecture & Media Information Gateway (ADAM). Disponível em: http://adam.ac.uk/. Acesso em: 20/07/2004.

Australian Institute of Health and Welfare Knowledgebase. Disponível em: http://www.aihw.gov.au/knowledgebase/. Acesso em: 23/07/2004.

Beyond Bookmarks. Disponível em http://www.iastate.edu/~CYBERSTACKS\CTW.htm. Acesso em: 27/07/04.

Cadê. Disponível em: http://www.cade.com.br . Acesso em: 15/04/2004.

Cataloging and Retrieval of Information Over Networks Aplications (Catriona II). Disponível em: http://catriona2.lib.strath.ac.uk/catriona/. Acesso: 28/07/04.

Comissão Brasileira de Bibliotecas Universitárias (CBBU). Disponível em: http://www.bczm.ufrn.br/cbbu/. Acesso em 10/07/2004.

DESIRE *Registry*. Disponível em: http://desire.ukoln.ac.uk/registry/. Acesso em: 23/07/04.

Development of a European Service for Information on Research and Education (DESIRE). Disponível em: http://www.desire.org/. Acesso em: 19/07/2004.

Development of a European Service for Information on Research and Education (DESIRE) *Registry*. Disponível em: http://desire.ukoln.ac.uk/registry/. Acesso em 23/07/04.

Dogpile. Disponível em http://www.dogpile.com/. Acesso em: 19/07/2004.

Dublin Core Metadata Element Set, Version 1.1: Reference Description. Disponível em: http://dublincore.org/documents/2003/02/04/dces. Acesso em: 28.07.04.

Dublin Core Metadata Initiative (DCMI). Disponível em: http://dublincore.org/. Acesso em: 28/07/04.

Dublin Core Metadata Initiative Usage Guide. Disponível em: http://dublincore.org/documents/usageguide/. Acesso em: 27/07/04.

Dublin Core Metadata Registry. Disponível em: http://dublincore.org/dcregistry/. Acesso em: 27/07/04.

Edinburgh Enginnering Virtual Library (EEVL). Disponível em: http://www.eevl.ac.uk/. Acesso em: 20/07/2004.

Encyclozine. Disponível em: http://encyclozine.com/Reference/Library/Classification/. Acesso em: 22/07/04.

Environmental Protection Agency. Disponível em: http://www.epa.gov/edr/. Acesso em: 23/07/04

Getty Institute Glossary. Disponível em: http://www.getty.edu/research/conducting_research/standards/intrometadata/4_glossary/index.html. Acesso em: 19/07/2004.

IFLA Functional Requirements for Bibliographic Records (FRBR). Disponível em: http://www.ifla.org/VII/s13/frbr/frbr.pdf. Acesso em: 27/07/04.

International Standard Bibliographic Description. (ISBD). Disponível em: http://www.infla.org/VI/3/nd1/isbdlist.htm. Acesso em: 01/08/2004.

Library of Congress. Disponível em: http://www.loc.gov/. Acesso em: 12/06/2004.

Mamma. Disponível em http://www.mamma.com. Acesso em: 19/07/2004.

MARC21. Disponível em: http://www.loc.gov/marc/. Acesso em: 27/07/04.

National Center for Computing Applications (NCSA). Disponível em: http://www.ncsa.uiuc.edu/. Acesso em: 13/07/2004.

Online Computer Library Center (OCLC). Disponível em: http://www.oclc.org/. Acesso em: 12/07/2004.

Online Dictionary for Information Science (ODLIS). Disponível em http://lu.com/odlis/odlis_r.cfm. Acesso em: 20/07/2004.

Páginas Brasileiras. Disponível em: http://www.prossiga.br/paginasbrasileiras. Acesso em: 20/07/2004.

Prossiga. Disponível em: http://www.prossiga.br. Acesso em: 20/07/2004.

Resource Organization and Subject-based Services (ROADS) *Registry*. Disponível: em http://www.ukoln.ac.uk/metadata/roads/templates/. Acesso em: 23/07/04.

Resource Organization and Subject-based Services (ROADS). Disponível em: http://www.ilrt.bris.ac.uk/roads/. Acesso em: 23/07/2004.

Savvy Search. Disponível em http://www.search.com/. Acesso em: 19/07/2004.

U. K. Office for Library and Information Networking (UKOLN). Disponível em http://www.ukoln.ac.uk/metadata/. Acesso em: 27/07/04.

Understanding Marc Bibliographic. Disponível em: *http://www.loc.gov/marc/umb*. Acesso em: 24/06/2004.

WebCrawler. Disponível em: http://www.*web*crawler.com/. Acesso em: 19/07/2004.

WorlCat. Disponível em: http://www.oclc.org/worldcat/. Acesso em: 22/07/2004.

World Wide Web Consortium (W3C). Disponível em: http://www.w3.org/. Acesso em: 19/07/2004.

World Wide Web Virtual Library. Disponível em: http://www.vlib.org. Acesso em: 15/04/2004.

Yahoo ! Disponível em: http://www.yahoo.com. Acesso em 19/07/2004.

ANEXO 1 - MAPEAMENTO DOS ESQUEMAS DE METADADOS NO EXTERIOR

METADADOS GERAIS

Diversas Comunidades

1. Dublin Core (DC)

Definição/Objetivo:
Os elementos de metadados do Dublin Core compõem um esquema para a descrição de recursos e sua aplicação é geral. Seu objetivo original é facilitar a descoberta de objetos de informação na *Web*.

Instituição responsável:
Dublin Core Metadata Initiative.

Comunidades atendidas:
Várias comunidades de diferentes áreas.

Homepage:
http://www.dublincore.org. (Acesso em: 27/07/2004).

URL para os elementos:
http://www.dublincore.org/documents/dces. (Acesso em: 27/07/2004).

METADADOS ESPECIALIZADOS POR ÁREA

Arte e Arquitetura

2. Categories for the Description of Works of Art (CDWA)

Definição/Objetivo:
O objetivo do CDWA é atingir um consenso na comunidade sobre os elementos básicos para a descrição de trabalhos de arte.

Instituições responsáveis:
Art Information Task Force (AITF), um projeto do *College Art Association of America* e do GETTY *Information Institute*.

Comunidades atendidas:
Comunidades que fornecem e utilizam informação sobre arte: historiadores de arte, curadores de museus, profissionais de recursos visuais, bibliotecários especializados em arte, administradores de informação e técnicos especialistas da área.

Homepage:
http://www.getty.edu/research/conducting_research/standards/cdwa. (Acesso em: 28/07/2004).

URL para os elementos:
http://www.getty.edu/research/conducting_research/standards/cdwa/4_categories/index.html. (Acesso em: 28/07/2004).

3. Visual Resources Association Core (VRA Core)

Definição/Objetivo:
O padrão VRA Core é desenhado para facilitar o compartilhamento de informações sobre trabalhos e imagens de coleções de recursos visuais.

Instituição responsável:
Visual Resources Association Data Standards Committee.

Comunidades atendidas:
Comunidades que fornecem e usam informação de arte: historiadores de arte, curadores de museus, profissionais de recursos visuais, bibliotecários especializados em arte, administradores de informação e técnicos especialistas da área.

Homepage:
http://vraweb.org/vracore3.htm. (Acesso em: 28/07/2004).

URL para os elementos:
http://vraweb.org/vracore3.htm#core. (Acesso em: 28/07/2004).

Biologia e Meio Ambiente

4. NBII biological metadata (NBII)

Definição/Objetivo:
O NBII é um programa colaborativo entre parceiros do âmbito federal, estadual e internacional, de cunho não-governamental, acadêmico e da indústria privada, para aumentar a acessibilidade aos dados e informações sobre recursos biológicos.

Instituições responsáveis:
Biological Data Working Group do *Federal Geographic Data Committee* (FGDC) e a *Biological Resources Division* do *United States Geological Survey* (USGS).

Comunidades atendidas:
Biólogos.

Homepage:
http://www.nbii.gov/datainfo/metadata/. (Acesso em: 30/07/2004).

URL para os elementos:
http://www.fgdc.gov/standards/documents/standards/biodata/biodatap.html. (Acesso em: 30/07/2004)

5. Ecological Metadata Language (EML)

Definição/Objetivo:
O objetivo do EML é descrever dados relevantes para a disciplina de ecologia.

Instituição responsável:
Knowledge Network for Biocomplexity (KNB).

Comunidades atendidas:
Ecologistas.

Homepage:
http://knb.ecoinformatics.org/software/eml/. (Acesso em: 30/07/2004).

URL para os elementos:
http://knb.ecoinformatics.org/software/eml/eml-2.0.0/index.html. (Acesso em: 30/07/2004).

6. Darwin Core

Definição Objetivo:
Descrever coleções de história natural e bases de dados de observação.

Instituições responsáveis:
Z39.50 Biology Implementors Group (CBIG), o projeto *Especies Analyst* e o *Natural History Museum and Biodiversity Research Center* da *Kansas University*.

Comunidades atendidas:
Pesquisadores de ciências naturais.

Homepage:
http://speciesanalyst.net/docs/dwc/index.html. (Acesso em: 29/07/2004).

URL para os elementos:
http://digir.net/schema/conceptual/darwin/2003/1.0/darwin2.xsd. (Acesso em: 29/07/2004).

Ciências Sociais

7. Data Documentation Initiative (DDI)

Definição/Objetivo:
Esforço internacional para estabelecer um padrão para a documentação técnica que descreve os dados das ciências sociais.

Instituições responsáveis:
International Association of Social Science Information Service and Technology (IASSIST), *Inter-university Consortium for Political and Social Research* (ICPSR), *Council of European Social Science Data Services* (CESSDA), *International Federation of Data Organizations* (IFDO), *Roper Center for Public Opinion Research.*

Comunidades atendidas:
Cientistas sociais.

Homepage:
http://www.icpsr.umich.edu/DDI/. (Acesso em: 29/07/2004).

URL para os elementos:
http://www.icpsr.umich.edu/DDI/users/dtd/index.html. (Acesso em: 29/07/2004).

Educação

8. Gateway to Educational Materials (GEM)

Definição/Objetivo:
O padrão GEM tem como objetivo oferecer acesso às coleçoes de materiais educacionais na *internet*, ainda não catalogados, diponíveis em *sites* de nstituições comerciais, federais e estaduais, de universidades e instituições não-lucrativas.

Instituição responsável:
United States Education Department.

Comunidades atendidas:
Profissionais de educação.

Homepage:
http://www.geminfo.org/index.html . (Acesso em: 29/07/2004).

URL para os elementos:
http://www.geminfo.org/Workbench/Metadata/index.html . (Acesso em: 29/07/2004).

9. Learning Object Metadata (LOM)

Definição/Objetivo:
O padrão LOM tem como objetivo descrever e gerenciar objetos para aprendizagem (entidades digitais ou não-digitais, usadas, re-utilizadas ou referenciadas durante o aprendizado, que utilizam algum instrumental tecnológico. Exemplos: ambientes de treinamento interativos, sistemas de aprendizado a distância.

Instituições responsáveis:
Institute of Electrical and Electronics Engineers (IEEE), C*omputer Society*, *Learning Technology Standards Committee.*

Comunidades atendidas:
Profissionais de educação.

Homepage:
http://ltsc.ieee.org/wg12/index.html. (Acesso em: 29/07/2004).

10. Sharable Content Object Reference Model (SCORM)

Definição/Objetivo:
O padrão SCORM tem como objetivo descrever e gerenciar objetos para aprendizagem (entidades digitais ou não-digitais, usadas, re-utilizadas ou referenciadas durante o aprendizado, que utilizam algum instrumental tecnológico. Exemplos: ambientes de treinamento interativos, sistemas de aprendizado a distância.

Instituição responsável:
Advanced Distributed Learning Network.

Comunidades atendidas:
Profissionais de educação.

Homepage:
http://www.adlnet.org/index.cfm?fuseaction=scormabt. (Acesso em: 29/07/2004).

URL para os elementos:
http://www.adlnet.org/index.cfm?fuseaction=SCORMDown . (Acesso em: 29/07/2004).

Geociências

11. Content Standard for Digital Geospatial Metadata (CSDGM)

Definição/Objetivo:
O objetivo do CSDGM é descrever recursos geo-espaciais digitais.

Instituição responsável:
Federal Geographic Data Committee (FGDC).

Comunidades atendidas:
Agências de dados geoespaciais do governo federal e do setor privado.

Homepage:
http://www.fgdc.gov/metadata/contstan.html . (Acesso em: 29/07/2004).

URL para os elementos:
http://www.fgdc.gov/metadata/csdgm/. (Acesso em: 29/07/2004).

Lingüística e Literatura

12. Text Encoding Initiative (TEI)

Definição/Objetivo:
O objetivo do TEI é desenvolver diretrizes para codificar textos lingüísticos e literários, utilizando a linguagem SGML e encorajar sua utilização e intercâmbio entre os bibliotecários, museólogos, editores e pesquisadores de ciências humanas.

Instituições responsáveis:
Association for Computers and the Humanities, *Association for Computational Linguistics* e *Association for Literary and Linguistic Computing*.

Comunidades atendidas:
Bibliotecários, museólogos, editores e universitários.

Homepage:
http://www.tei-c.org. (Acesso em: 27/07/2004).

URL para os elementos:
http://www.tei-c.org/Guidelines2/index.html. (Acesso em: 27/07/2004).

APLICAÇÕES DE METADADOS

Arquivos

13. Encoded Archival Description (EAD)

Definição/Objetivo:
O EAD é um padrão para codificar guia de arquivo usando SGML.
Instituições responsáveis:
Library of Congress em parceria com a *Society of American Archivists*.
Comunidades atendidas:
Arquivos e repositórios de manuscritos.
Homepage:
http://www.loc.gov/ead. (Acesso em: 28/07/2004).
URL para os elementos:
http://www.loc.gov/ead/tglib/index.html. (Acesso em: 28/07/2004).

Bibliotecas

14. Machine Readable Catalog (MARC21)

Definição/Objetivo:
O padrão MARC21 objetiva descrever informação bibliográfica.
Instituição responsável:
Library of Congress.
Comunidades atendidas:
Bibliotecas.
Homepage:
http://www.loc.gov/marc/. (Acesso em: 28/07/2004).
URL para os elementos:
http://www.loc.gov/marc/bibliographic/ecbdhome.html. (Acesso em: 28/07/2004).

Comércio de Livros

15. Onix International

Definição/Objetivo:
Desenvolvido por editores para troca de informação comercial en forma eletrônica entre vendedores, distribuidores e outras partes da cadeia de distribuição de livros.
Instituição responsável:
Advanced Distributed Learning Network.
Comunidades atendidas:
Editores, vendedores e distribuidores de livros.
Homepage:
http://www.editeur.org/. (Acesso em: 29/07/2004).
URL para os elementos:
http://www.editeur.org/ONIX_Code_Lists_Issue_2.PDF . (Acesso em: 29/07/2004).

Direitos Autorais

16. Interoperability of Data in E-Commerce Systems (INDECS)

Definição/Objetivo:
Modelo semântico para descrever a propriedade intelectual.
Instituição responsável:
Indecs Framework Ltd.
Comunidades atendidas:
Produtoras de filmes, produtoras de música, gravadoras, editoras de livros e revistas.
Homepage:
http://www.indecs.org/pdf/framework.pdf. (Acesso em: 29/07/2004).
URL para os elementos:
http://www.indecs.org/pdf/framework.pdf. (Acesso em: 29/07/2004).

17. Open Digital Rights Language (ODRL)

Definição/Objetivo:
A ODRL é uma linguagem que visa padronizar a descrição de direitos autorais sobre recursos eletrônicos.

Instituição responsável:
IPR Systems Pty Ltd.

Comunidades atendidas:
Comunidades que administram os direitos autorais sobre recursos digitais.

Homepage:
http://odrl.net/. (Acesso em: 29/07/2004).

URL para os elementos:
http://www.w3.org/TR/odrl/. (Acesso em: 29/07/2004).
http://odrl.net/1.1/ODRL-EX-11-DOC/index.html . (Acesso em: 29/07/2004).

18. eXtensible rights Markup Language (XRML)

Definição Objetivo:
XrML é uma linguagem para especificar e administrar de forma segura informação de direitos autorais sobre recursos digitais e serviços.

Instituição responsável:
ContentGuard.

Comunidades atendidas:
Comunidades que administram os direitos autorais sobre recursos digitais.

Homepage:
http://www.xrml.org/. (Acesso em: 29/07/2004).

URL para os elementos:
http://www.xrml.org/Reference/XrMLTechnicalOverviewV1.pdf. (Acesso em: 29/07/2004).

Informação do Governo

19. Goverment Information Locator Service (GILS)

Definição/Objetivo:
O padrão GILS tem como objetivo identificar, localizar e descrever recursos de informação do Governo Federal americano, incluindo recursos de informação eletrônicos.

Instituição responsável:
United States Government.

Comunidades atendidas:
Agências Governamentais.

Homepage:
http://www.gils.net (Acesso em: 28/07/2004).

URL para os elementos:
http://www.gils.net/prof_v2.html#sec_8 . (Acesso em: 28/07/2004).

Metadados Estruturados

20. Exchange Format For Electronic Components and Texts (EFFECT)

Definição/Objetivo:
O objetivo do EFFECT é dar suporte ao processo de distribuição de periódicos e artigos electrônicos desde as editoras até as bibliotecas ou outras organizações.

Instituição responsável:
Elsevier Science.

Comunidades atendidas:
Editores, vendedores, distribuidores de periódicos e artigos eletrônicos.

Homepage:
http://support.sciencedirect.com/sdos/effect41.pdf . (Acesso em: 29/07/2004)

URL para os elementos:
http://support.sciencedirect.com/sdos/effect41.pdf . (Acesso em: 29/07/2004).

21. Berkeley Electronic Binding Project (EBIND)

Definição/ bjetivo:
O objetivo do EBIND é descrever metadados estruturais para recursos digitalizados em forma de imagens.

Instituição responsável:
University of California e *Berkeley Library*.

Comunidades atendidas:
Bibliotecários e arquivistas.

Homepage:
http://sunsite3.berkeley.edu/Ebind/. (Acesso em: 29/07/2004).

URL para os elementos:
http://sunsite.berkeley.edu/Ebind/ebind.dtd. (Acesso em: 29/07/2004).

22. Making of America II (MOA2)

Definição/Objetivo:
Padrão para codificar metadados descritivos, administrativos e estruturais, junto com o conteúdo dos recursos digitalizados.

Instituições responsáveis:
Integrantes da *Digital Library Federation*: *University of California*, *Berkeley Library*, *Cornell University*, *New York Public Library*, *Pennsylvania State University* e *Standford University*.

Comunidades atendidas:
Bibliotecários e arquivistas.

Homepage:
http://sunsite3.berkeley.edu/MOA2/. (Acesso em: 29/07/2004).

URL para os elementos:
http://sunsite.berkeley.edu/moa2/papers/moa2dtd2.htm. (Acesso em: 29/07/2004).

23. MPEG-7 (ISO/IEC 15938)

Definição/Objetivo:
Contém metadados descritivos, administrativos e estruturais para recursos de vídeo e áudio.

Instituição responsável:
Motion Picture Experts Group.

Comunidades atendidas:
Produtores de vídeo e áudio digital.

Homepage:
http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm. (Acesso em: 29/07/2004).

URL para os elementos:
http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm. (Acesso em: 29/07/2004).

Preservação

24. Open Archival Information System (OAIS)

Definição/Objetivo:
O objetivo do OAIS é descrever a infraestrutura de informação quedá suporte ao processo de preservaçao digital.

Instituições responsáveis:
Online Computer Library Center (OCLC), *Research Libraries Group* (RLG) e o *Working Group on Preservation Metadata*.

Comunidades atendidas:
Comunidades engajadas na preservação digital de recursos.

Homepage:
http://www.oclc.org/research/projects/pmwg/pm_framework.pdf . (Acesso em: 29/07/2004).

URL para os elementos:
http://www.oclc.org/research/projects/pmwg/pm_framework.pdf. (Acesso em: 29/07/2004).

Tecnologia de Áudio e Vídeo

25. Technical Metadata for Digital Still Images (TMDSI)

Definição/Objetivo:
Este padrão objetiva facilitar o desenvolvimento de aplicações para validação, gerenciamento, migração e o processamento de imagens de valor permanente.

Instituições responsáveis:
National Information Standards Organization e *AIIM International.*

Comunidades atendidas:
Instituições culturais, editores e outras organizações engajadas na digitalização de materiais visuais pertencentes a coleções históricas.

Homepage:
http://www.niso.org/standards/resources/Z39_87_trial_use.pdf. (Acesso em: 29/07/2004)

URL para os elementos:
http://www.niso.org/standards/resources/Z39_87_trial_use.pdf. (Acesso em: 29/07/2004).

26. Audio Technical Metadata Extension Schema (AUDIOMD)

Definição/Objetivo:
O objetivo do AUDIOMD é descrever arquivos de áudio digital e sua fonte digital ou analógica.

Instituição responsável:
Library of Congress.

Comunidades atendidas:
Profissionais na área de audio digital.

Homepage:
http://lcweb.loc.gov/rr/mopic/avprot/metsmenu2.html. (Acesso em: 29/07/2004).

URL para os elementos:
http://www.loc.gov/rr/mopic/avprot/DD_AMD.html. (Acesso em: 29/07/2004).

27. Video Technical Metadata Extension Schema (VIDEOMD)

Definição/objetivo:
O objetivo do VIDEOMD é descrever arquivos de vídeo digital e sua fonte digital ou analógica.

Instituição responsável:
Library of Congress.

Comunidades atendidas:
Profissionais na área de vídeo digital.

Homepage:
http://lcweb.loc.gov/rr/mopic/avprot/metsmenu2.html. (Acesso em: 29/07/2004).

URL para os elementos:
http://lcweb.loc.gov/rr/mopic/avprot/DD_VMD.html. (Acesso em: 29/07/2004).

ANEXO 2 - QUESTIONÁRIO PARA COLETA DE DADOS

1. Tem conhecimento sobre metadados?

   Sim ( ) Não ( )

1.1. Em caso positivo, indicar um conceito ou definição, extraído da literatura e que corresponda às suas idéias.

1.2. Indique o(s) esquema(s) de metadados ("metadata scheme") que conhece:

   Dublin Core ( )
   IEEELOM ( )
   MARC ( )
   IMS ( )
   GILS ( )
   Outros (especificar):

2. No seu trabalho, utiliza metadados?

   Sim ( ) Não ( )

2.1. Justifique a resposta, em caso negativo ou positivo.

2.2. Em caso positivo, qual o esquema de metadados ("metadata scheme") utilizado e em que tipo de serviço de informação:

3. Nome completo:

   Cargo ou função:

   Departamento/Faculdade/Instituto:

   Instituição superior a qual é vinculado :

ANEXO 3 - INFORMAÇÕES SOBRE AS INSTITUIÇÕES PESQUISADAS

REGIÃO NORTE (3 bibliotecas)

Amazonas :

1. **UFAM. Universidade Federal do Amazonas**
   URL: http://biblioteca.ufam.edu.br/biblioteca/php/opcoes.php
   e-mails: dsa.bc@ufam.edu.br, diretor-bc@ufam.edu.br

Pará :

2. **UFPA. Universidade Federal do Pará**
   URL: http://www.ufpa.br/bc/
   e-mail: bitar@ufpa.br

3. **UEPA. Universidade do Estado do Pará**
   URL: http://www.uepa.br/bib/
   e-mail: bibuepa@uepa.br

REGIÃO NORDESTE (9 bibliotecas)

Bahia :

4. **UNEB. Universidade Estadual da Bahia**
   URL: http://www.bib.uneb.br/
   e-mail: bcentral@uneb.br

5. **UFBA. Universidade Federal da Bahia**
   URL: http://www.bib.ufba.br/ufba/
   e-mail: bcdir@ufba.br

Maranhão:

6. **UFMA. Universidade Federal do Maranhão**
   URL: http://www.biblioteca.ufma.br/
   e-mail: ufmabc@ufma.br

Pernambuco:

7. **Universidade Federal de Pernambuco**
   URL: http://www.ufpe.br/sib/
   e-mail: bcufpe@npd.ufpe.br

Rio Grande do Norte:

8. Universidade Federal do Rio Grande do Norte
URL: http://www.bczm.ufrn.br/
e-mail: bcdir@bczm.ufrn.br

Sergipe:

9. UFS. Universidade Federal de Sergipe
URL: http://www.biblioteca.ufs.br/
e-mail: **bicen@ufs.br**

Alagoas:

10. UFAL. Universidade Federal de Alagoas
URL: http://www.sibi.ufal.br/
e-mails: helena@sibi.ufal.br, bibcentral@sibi.ufal.br

Ceará:

11. UECE. Universidade Estadual do Ceará
URL: http://www.uece.br/biblioteca/
e-mails: ampb@uece.br, bibliot@uece.br

12. UFC. Universidade Federal do Ceará
URL: http://www.biblioteca.ufc.br/
e-mail: bu@ufc.br

REGIÃO CENTRO-OESTE (4 bibliotecas)

Distrito Federal

13. UCB. Universidade Católica de Brasília
URL: http://www.biblioteca.ucb.br/BC.htm
e-mail: bcentral@ucb.br

14. UNB. Universidade de Brasília
URL: http://www.bce.unb.br/
e-mail: direcao@bce.unb.br

Goiás:

15. UFG. Universidade Federal de Goiás
URL: http://www.bc.ufg.br/
e-mails: valeria@bc.ufg.br , direcao@bc.ufg.br

Mato Grosso do Sul:

16. **UFMS. Universidade Federal do Mato Grosso do Sul**
    URL: http://www.cbc.ufms.br/
    e-mail: lucia@nin.ufms.br

REGIÃO SUDESTE (15 bibliotecas)

Espírito Santo:

17. **UFES. Universidade Federal do Espírito Santo**
    URL: http://www.bc.ufes.br/index.htm
    e-mail: direcao@bc.ufes.br

Minas Gerais:

18. **PUC-Minas. Pontifícia Universidade Católica de Minas Gerais**
    URL: http://www.pucminas.br
    e-mail: bibdir@pucminas.br

19. **UEMG. Universidade do Estado de Minas Gerais**
    URL: http://www.uemg.br
    e-mail: biblioteca@ituiutaba.uemg.br

20. **UFMG. Universidade Federal de Minas Gerais**
    URL: http://www.bu.ufmg.br
    e-mail: dir@bu.ufmg.br

21. **UFOP. Universidade Federal de Ouro Preto**
    URL: http://www.sisbin.ufop.br/
    e-mail: sisbin@biblio.ufop.br

22. **UFU. Universidade Federal de Uberlândia**
    URL: http://www.bibliotecas.ufu.br/
    e-mail: elzac@dirbi.ufu.br

23. **UFV. Universidade Federal de Viçosa**
    URL: http://www.ufv.br/bbt
    e-mail: masoares@ufv.br

Rio de Janeiro:

24. **PUC-Rio. Pontifícia Universidade Católica do Rio de Janeiro**
    URL: http://www.dbd.puc-rio.br/
    e-mail: anamaria@dbd.puc-rio.br

25. **UFF. Universidade Federal Fluminense**
    URL: http://www.ndc.uff.br/
    e-mail: sir@ndc.uff.br

26. **UERJ. Universidade Estadual do Rio de Janeiro**
    URL: http://www2.uerj.br/~rsirius/
    e-mail: rsirius@uerj.br

27. **Uni-Rio. Universidade do Rio de Janeiro**
    URL: http://www.unirio.br/biblioteca/bibliotecas.htm
    e-mail: bp_direcao@unirio.br

28. **UFRJ. Universidade Federal do Rio de Janeiro**
    URL: http://www.sibi.ufrj.br
    e-mail: crefdir@sibi.ufrj.br

São Paulo:

29. **PUC-SP. Pontifícia Universidade Católica de São Paulo**
    URL: http://biblio.pucsp.br
    e-mail: biblio@pucsp.br

30. **USP. Universidade de São Paulo**
    URL: http://www.usp.br/sibi/
    e-mail: dtsibi@org.usp.br

31. **UNICAMP. Universidade Estadual de Campinas**
    URL: http://www.unicamp.br/bc/
    e-mail: bibcen@obelix.unicamp.br

REGIÃO SUL (4 bibliotecas)

Paraná:

32. **PUC-PR. Pontifícia Universidade Católica de Paraná**
    URL: http://www.biblioteca.pucpr.br/
    e-mail: heloisa.anzolin@pucpr.br

Rio Grande do Sul:

33. **UFPEL. Universidade Federal de Pelotas**
    URL: http://www.ufpel.tche.br/prg/sisbi/
    e-mail: zibetti@ufpel.edu.br

34. **UERGS. Universidade Estadual do Rio Grande do Sul**
    URL: http://www.uergs.rs.gov.br/interno/setores/biblio.htm
    e-mail: biblioteca@uergs.rs.gov.br

35. **UFSC. Universidade Federal de Santa Catarina**
    URL: http://www.bu.ufsc.br/
    e-mail: sigrid@bu.ufsc.br