UNIVERSIDADE FEDERAL DE MINAS GERAIS
ESCOLA DE CIÊNCIA DA INFORMAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO

A IMPORTÂNCIA DOS METADADOS EM BIBLIOTECAS DIGITAIS:

da organização à recuperação da informação

Eduardo Ribeiro Felipe

Belo Horizonte

2012

Eduardo Ribeiro Felipe

A IMPORTÂNCIA DOS METADADOS EM BIBLIOTECAS DIGITAIS:
da organização à recuperação da informação

Dissertação apresentada ao Programa de Pós Graduação da Escola de Ciência da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Mestre em Ciência da Informação.

Área de concentração: Produção, Organização e Utilização da Informação.
Linha de pesquisa: Organização e Uso da Informação.

Orientadora: Prof. Dra. Gercina Ângela de Oliveira Lima

Belo Horizonte

2012

Felipe, Eduardo Ribeiro.

F315i A importância dos metadados em bibliotecas digitais [manuscrito] : da organização à recuperação da informação / Eduardo Ribeiro Felipe. – 2012.
110 f. : il., enc.

Orientadora: Gercina Ângela de Oliveira Lima.
Dissertação (mestrado) – Universidade Federal de Minas Gerais, Escola de Ciência da Informação.
Referências: f. 94-100
Apêndices: f. 101-105
Anexos: f. 106-110

1. Ciência da informação – Teses. 2. Sistemas de recuperação da informação – Teses. 3. Bibliotecas digitais – Teses. 4. Metadados – Teses. I. Título. II. Lima, Gercina Ângela Borém de Oliveira. III. Universidade Federal de Minas Gerais, Escola de Ciência da Informação.

CDU: 02:004

Ficha catalográfica: Biblioteca Profª Etelvina Lima, Escola de Ciência da Informação da UFMG

UFMG

Universidade Federal de Minas Gerais
Escola de Ciência da Informação
Programa de Pós-Graduação em Ciência da Informação

FOLHA DE APROVAÇÃO

"A IMPORTÂNCIA DOS METADADOS EM BIBLIOTECAS DIGITAIS: DA ORGANIZAÇÃO À RECUPERAÇÃO DA INFORMAÇÃO"

Eduardo Ribeiro Felipe

Dissertação submetida à Banca Examinadora, designada pelo Colegiado do Programa de Pós-Graduação em Ciência da Informação da Universidade Federal de Minas Gerais, como parte dos requisitos à obtenção do título de "**Mestre em Ciência da Informação**", Linha de Pesquisa: "**Organização e Uso da Informação - OUI**".

Dissertação aprovada em: 20 de dezembro de 2012.

Por:

Profa. Dra. Gercina Ângela Borém de Oliveira Lima - ECI/UFMG (Orientadora)

Profa. Dra. Madalena Martins Lopes Naves - Profa. Aposentada - ECI/UFMG

Prof. Dr. Maurício Barcellos Almeida - ECI/UFMG

Profa. Dra. Renata Maria Abrantes Baracho Porto - ECI/UFMG

Aprovada pelo Colegiado do PPGCI

Profa. Renata Maria Abrantes Baracho Porto
Coordenadora

Versão final Aprovada por

Profa. Gercina Ângela Borém de Oliveira Lima
Orientadora

Av. Antônio Carlos, 6627 – Sala 2003 – Campus Pampulha – Cx. Postal 1606 – CEP: 30161-970 – Belo Horizonte – MG
Telefone: (31)3409-6103 – Fax: (31)3409-5207 – www.eci.ufmg.br/ppgci – e-mail: ppgci@eci.ufmg.br

UFMG

Universidade Federal de Minas Gerais
Escola de Ciência da Informação
Programa de Pós-Graduação em Ciência da Informação

ATA DA DEFESA DE DISSERTAÇÃO DE **EDUARDO RIBEIRO FELIPE**, matrícula: 2008651333

Às 14:00 horas do dia 20 de dezembro de 2012, reuniu-se na Escola de Ciência da Informação da UFMG a Comissão Examinadora aprovada pelo Colegiado do Programa de Pós-Graduação em Ciência da Informação em 04/12/2012, para julgar, em exame final, o trabalho intitulado ***A importância dos metadados em bibliotecas digitais: da organização à recuperação da informação***, requisito final para obtenção do Grau de MESTRE em CIÊNCIA DA INFORMAÇÃO, Área de Concentração: Produção, Organização e Utilização da Informação, Linha de Pesquisa: Organização e Uso da Informação - OUI. Abrindo a sessão, a Presidente da Comissão, Profa. Dra. Gercina Ângela Borém de Oliveira Lima, após dar conhecimento aos presentes do teor das Normas Regulamentares do Trabalho Final, passou a palavra ao candidato para apresentação de seu trabalho. Seguiu-se a argüição pelos examinadores com a respectiva defesa do candidato. Logo após, a Comissão se reuniu sem a presença do candidato e do público, para julgamento e expedição do resultado final. Foram atribuídas as seguintes indicações:

Profa. Dra. Gercina Ângela Borém de Oliveira Lima – Orientadora	APROVADO
Profa. Dra. Madalena Martins Lopes Naves	APROVADO
Prof. Dr. Maurício Barcellos Almeida	APROVADO
Profa. Dra. Renata Maria Abrantes Baracho Porto	APROVADO

Pelas indicações, o candidato foi considerado APROVADO.

O resultado final foi comunicado publicamente ao candidato pela Presidente da Comissão. Nada mais havendo a tratar, a Presidente encerrou a sessão, da qual foi lavrada a presente ATA que será assinada por todos os membros participantes da Comissão Examinadora.

Belo Horizonte, 20 de dezembro de 2012

Profa. Dra. Gercina Ângela Borém de Oliveira Lima
ECI/UFMG (Orientadora)

Profa. Dra. Madalena Martins Lopes Naves
Profa. Aposentada - ECI/UFMG

Prof. Dr. Maurício Barcellos Almeida
ECI/UFMG

Profa. Dra. Renata Maria Abrantes Baracho Porto
ECI/UFMG

Obs: Este documento não terá validade sem a assinatura e carimbo da Coordenadora.

Profa. Renata Maria Abrantes Baracho Porto
Coordenadora do Programa de Pós-Graduação em Ciência da Informação - ECI / UFMG

A Deus, acima de tudo, sempre.

A meus pais, Antônio e Julieta.

E a Renata e Luciana.

AGRADECIMENTOS

Sou muito grato a Deus por me conceder mais este momento de conquista em minha vida. Sou grato a Jesus por ter feito TUDO por mim e minha família, pela nossa saúde e prosperidade, grato por participar deste mistério: “Cristo em vós, esperança da glória.”

Um agradecimento especial à minha orientadora, Profa. Dra. Gercina Lima, pela forma como conduziu este trabalho e pela paciência e sensibilidade nas minhas dificuldades pelo caminho, que não foram poucas. A senhora tem mais que meu reconhecimento, tem o meu carinho, respeito e admiração. Muito obrigado por tudo.

À professora Madalena Naves, pela atenção e considerações que me ajudaram nesta caminhada.

Agradeço à minha família, aos meus pais, que batalharam muito para me educar e fazer tudo o que podiam para minha formação pessoal. As conquistas que posso colher hoje são fruto do esforço de vocês. Recebam minha eterna gratidão por tudo que fizeram e ainda fazem por mim.

À Luciana, pelo apoio e pelos acertos no caminho.

À Renata, pela torcida, apoio e amor de família.

Ao meu amigo Edson Tadeu, que há anos foi o responsável por me apresentar a primeira ferramenta de programação orientada a objetos e acabou por mostrar-se mais que um professor, tornou-se um grande amigo que tem me ajudado até hoje. Obrigado pelo apoio no protótipo, pelas orientações de grande valor e pela amizade imensurável.

À minha tia Adélia, pela torcida e orações constantes.

Aos meus tios Carlos e Lécia por me apoiar tanto, até quando eu mesmo não acreditava. Pelos momentos inesquecíveis “no sítio” e também por “aceitar meus gatos”, eu sei que dou trabalho, mas também sei que sou amado.

Aos professores da ECI/UFMG, que possibilitaram um ensino de muita qualidade. Este trabalho tem um pouco de cada um de vocês.

À Sônia e à Cláudia do NITEG, muito obrigado pelo apoio.

À Gisele e à Nely da secretaria do PPGCI, pelo apoio e orientações.

Aos professores Jorge Tadeu e Marcelo Bax, pela confiança e pelas oportunidades em lecionar na pós-graduação.

Ao professor Maurício Barcellos e à professora Renata Baracho, pelo convite aceito.

Ao professor Dagobert Soergel, pelas opiniões que contribuíram na interface de cadastro dos Metadados e demais considerações sobre o protótipo.

Aos colegas do grupo MHTX, pelas discussões enriquecedoras.

À especialista Maria Helena, pelas aulas preciosas no formato Marc.

À D. Glória Amorim, pelas importantes orientações.

Ao Dino e à Cléo, *in memoriam*, e ao Nico e à Nina, eles sabem por que.

“E, demais disto, filho meu, atenta: não há limite para fazer livros,
e o muito estudar é enfado da carne.”
Eclesiastes, 12:12

RESUMO

As Bibliotecas sofreram uma grande mudança a partir da invenção do computador. Seus processos de organização, manutenção, pesquisa e recuperação precisaram ser revistos à luz da capacidade virtual deste novo ambiente de representação da informação. Este trabalho realiza uma investigação sobre a capacidade dos metadados em contribuir com o ambiente da Biblioteca Digital: (a) possibilitando a definição de descritores para identificação de um documento; (b) estabelecendo um padrão para gravação e recuperação das informações relativas a um item específico; (c) permitindo sua recuperação através de mecanismos padronizados para o acesso ao documento. Um protótipo de *software* foi desenvolvido a partir da pesquisa, evidenciando a capacidade desse mecanismo de organizar e permitir a recuperação de um item bibliográfico eletrônico submetido ao padrão selecionado. A fundamentação teórico-metodológica teve como princípios as bibliotecas digitais, metadados e padrões de organização de metadados. As conclusões direcionam à valorização dessas estruturas de dados alinhadas a uma política informacional, possibilitando um acervo coeso e de fácil recuperação.

Palavras-chaves: Produção, Organização e Utilização da Informação, Metadados, Biblioteca Digital, Documento Eletrônico, Taxonomia, Protótipo.

ABSTRACT

Libraries have experienced a great change since the invention of the computer. The organization processes, maintenance, and search and recovery operations needed to be revised, in light of the new capacity of this new virtual capacity environment of information representation. This paper reports an investigation into the ability of metadata to contribute to the Digital Library's environment: (a) allowing the definition of descriptors to identify a document; (b) establishing a standard for recording and retrieving information related to a specific item; (c) allowing recovery of documents through standard mechanisms for access. A prototype software was developed from this research, to show the ability of this mechanism to organize and allow retrieval of an electronic bibliographic item submitted to a selected pattern. The theoretical and methodological groundings had the digital libraries, metadata and organization of metadata standards as the principles. As a result of this research, the software implementation was successful in its proposal and was capable of performing the processes investigated in the methodology. The conclusions pointed to the value of these data structures, when aligned to an informational politics, enabling a cohesive collection and of easy retrieval access.

Keywords: Production, Organization and Use of Information, Metadata, Digital Library, Electronic Document, Taxonomy, Prototype.

LISTA DE FIGURAS

Figura 1 – Esquema do IBICT e instituições de ensino 29
Figura 2 - Biblioteca Digital da Universidade Federal de Minas Gerais 30
Figura 3 - A função da elaboração de índices e resumos no quadro mais amplo da recuperação da informação 33
Figura 4 – Classificação Decimal de Dewey 37
Figura 5 - Tesauro de Folclore e Cultura Popular Brasileira 38
Figura 6 – Exemplo de ficha catalográfica 42
Figura 7 - Exemplo de informações em formato MARC 46
Figura 8 - Processamento de imagem com metadados 48
Figura 9 - *Software* de atribuição de metadados em imagens 49
Figura 10 – Modelo adotado pelo IBICT na biblioteca de teses e dissertações 52
Figura 11 – Esquema geral de funcionamento do modelo IBICT 54
Figura 12 - Documento XML bem formado 58
Figura 13 - A composição de uma tripla RDF 59
Figura 14 - Validador de código RDF e sua saída gráfica 61
Figura 15 - *Software* para visualização gráfica de arquivos RDF 62
Figura 16 - Comparativo das abordagens computacional e humana em uma página Web 63
Figura 17 - Macro processos para o protótipo 67
Figura 18 – Programa que permite traduzir os textos (*Strings*) com sua referência para o código fonte 69
Figura 19 – Janela inicial para configuração do idioma a ser usado 70
Figura 20 – Interface para permitir a mudança de idioma 70
Figura 21 - Configurações de pastas 71
Figura 22 - Configuração da taxonomia 73
Figura23 - Abertura de um documento PDF dentro do protótipo 76
Figura 24 - Extração dos termos baseados na taxonomia 78
Figura 25 - Preenchimento dos metadados 80
Figura 26 - Listagem completa dos documentos cadastrados 82
Figura 27 - Visualização do documento na interface de recuperação 83
Figura 28 - Recuperação com base na Taxonomia 84
Figura 29 - Recuperação com base em dois termos da Taxonomia 84
Figura 30 - Interface de recuperação por Linguagem Natural 85
Figura 31 - Resultado de pesquisa em Linguagem Natural com o termo “Gercina” 86

LISTA DE ABREVIATURAS

ANSI	American National Standards Institute
BT	Broader Term
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CC	Ciência da Computação
CI	Ciência da Informação
BDTD	Biblioteca Digital de Teses e Dissertações
DC	Dublin Core
DCMI	Dublin Core Metadata Initiative
ETD-MS	Electronic Theses and Dissertations
HTML	Hypertext Markup Language
IBICT	Instituto Brasileiro de Ciência e Tecnologia
IBM	International Business Machines
ID3	Metadata container
LC	Library of Congress
MARC	Machine Readable Cataloging Record
MDI	Multiple Document Interface
MHTX	Mapa Hipertextual
MP3	MPEG Audio Layer III
MPEG	Moving Picture Experts Group
MTD-BR	Padrão Brasileiro de Metadados de Teses e Dissertações
NDLTD	Networked Digital Library of Thesis and Dissertation
NISO	National Information Standards Organization
NT	Narrower Term
OAI-PMH	Protocolo aberto de arquivos para coleta de metadados
OCLC	Online Computer Library Center
OCR	Optical Character Recognition
PDF	Portable Document Format
PPGCI	Programa de Pós-Graduação em Ciência da Informação
RDF	Resource Definition Framework
SGML	Standart Generalized Markup Language
SRI	Sistema de Recuperação da Informação
UFMG	Universidade Federal de Minas Gerais
XMP	Extensible Metadata Plataform
XML	Extensible Markup Language
W3C	World Wide Web Consortium

SUMÁRIO

1 INTRODUÇÃO 15

1.1 Justificativa 16

1.2 Problema da pesquisa 18

1.3 Objetivos 18

1.3.1 Objetivo geral 19
1.3.2 Objetivos específicos 19

1.4 Estrutura do trabalho 19

2 FUNDAMENTAÇÃO TEÓRICO-METODOLÓGICA 20

2.1 Fundamentos 20

2.2 Bibliotecas digitais 23

2.2.1 Biblioteca Digital de Teses e Dissertações - IBICT 28

2.3 Indexação 31

2.4 Linguagens 36

2.5 Taxonomias 39

2.6 Metadados, padrões e linguagem 41

2.6.1 O que são os metadados 41
2.6.2 Machine-Readable Cataloging - MARC 21 44
2.6.3 Dublin Core 47
2.6.4 Biblioteca Digital de Teses e Dissertações – padrão MTD-BR 51

2.7 Esquemas para gravação de metadados 56

2.7.1 eXtensible Markup Language – XML 56
2.7.2 Resource Definition Framework – RDF 58
3 METODOLOGIA 64

3.1 Tipo de pesquisa 64

3.2 Universo de estudo 64

3.3 Projeção inicial para construção do protótipo MeTa 64

4 O DESENVOLVIMENTO DO PROTÓTIPO 68

4.1 Estruturas / Interface 68

4.2 Multilinguagem 69

4.3 Configuração 70

4.4.1 Configurações de Pastas 71

4.4.2 Definições / Seleção da taxonomia 72

4.5 Importação do documento 74

4.5.1 Documentos em formato proprietário 74

4.5.2 Documentos baseados em texto ou imagens 76

4.5.3 Documentos que possuem *copyright* 77

4.6 Identificador 77

4.7 Extração automática de termos baseados na taxonomia 77

4.8 Gravação de metadados 79

4.9 Recuperação da informação no protótipo 81

4.9.1 Pesquisa através da Listagem Completa dos itens 82

4.9.2 Pesquisa com base na Taxonomia 83

4.9.3 Pesquisa através da Linguagem Natural 84

5 CONSIDERAÇÕES E RECOMENDAÇÕES 87
REFERÊNCIAS 94
APÊNDICES 101

Apêndice A – Elementos do Dubin Core 101

ANEXOS 106

Anexo A – Taxonomia da Ciência da Informação 106

1 INTRODUÇÃO

Com o advento da informática, a criação de documentos eletrônicos foi facilitada de maneira significativa. Essa facilitação iniciou um processo de geração e publicação de informações jamais visto em toda a história da humanidade. Esse grande volume informacional cresce exponencialmente e oferece, da mesma forma, um alto grau de complexidade para o seu gerenciamento. Mecanismos de busca e de armazenamento estão em constante desenvolvimento, a fim de acompanhar essa escala, que é, hoje, praticamente imensurável. Nesse contexto, pesquisas específicas devem ser realizadas a fim de direcionar esforços para a gestão efetiva dessa enorme massa informacional. A Ciência da Informação deve, também, direcionar seus esforços nesse sentido, indo ao encontro do proposto por Saracevic:

> A Ciência da Informação é um campo dedicado às questões científicas e à prática profissional voltadas para os problemas da efetiva comunicação do conhecimento e de seus registros entre os seres humanos, no contexto social, institucional ou individual do uso e das necessidades de informação. No tratamento destas questões são consideradas de particular interesse as vantagens das modernas tecnologias informacionais. (SARACEVIC, 1991, p. 7).

A afirmação de Saracevic direciona a Ciência da Informação (CI) para uma abordagem voltada para a natureza da informação e sua comunicação com o ser humano. Por outro lado, a Ciência da Computação (CC) propõe algoritmos que manipulam essa informação. Bush (1945), em seu clássico “As we may think”, já vislumbrava essa complexidade e fazia uma previsão de que máquinas precisariam trabalhar a informação, e que elas o fariam com base no processo mental humano, o processo de associação. É nesse processo que Bush enfoca o uso da CI na abordagem de processos da comunicação humana, inspirando as modernas Bibliotecas Digitais.[1]

O tratamento da informação torna-se ainda mais desafiador na medida em que se vive um momento histórico em que informação, dado e conhecimento são tratados de forma diferenciada, seja nas organizações, seja nas instituições de ensino. A atenção, outrora direcionada à construção de mecanismos de criação e massificação da informação, volta-se agora para a complexa tarefa de lidar com quantidade sem detrimento da qualidade informacional. Nesse momento, a organização e a recuperação da informação constituem um grande desafio que pode ser compreendido pelos processos que abordam criação, indexação, armazenamento e disponibilização (publicação) da informação. Essa realidade informacional encontra-se evidenciada nos mais diversos mecanismos de gestão da

[1] Conceito a ser dissertado no Capítulo -- Bibliotecas Digitais.

informação, seja em grandes ou pequenas instituições. Este trabalho, inserido no contexto acadêmico, busca investigar o potencial dos metadados como contribuição aos processos informacionais. Nesse ambiente de tratamento informacional (profissional ou acadêmico), os profissionais responsáveis pela organização e disponibilização do acervo precisam lidar com uma enorme demanda de itens a serem inseridos no conjunto literário, seja em formato físico, seja em formato eletrônico. Segundo Gonçalves (2008), os mecanismos de organização e acesso à informação devem ainda ser mapeados de acordo com o perfil dos usuários, de modo a aumentar a exposição da informação documentária.

Acredita-se que a explosão informacional, iniciada na década de 60, criou um problema relacionado ao modo de trabalho desses profissionais da informação. O que se fazia no passado, através de procedimentos manuais, não mais é possível se realizar hoje. Percebe-se que, atualmente, cada área específica do conhecimento possui uma complexidade em seu tratamento, além do grande volume de material produzido cujas informações é necessário organizar, indexar e disponibilizar, em mecanismos de acesso. Esse modelo exige uma dinâmica automatizada ou, no mínimo, semiautomatizada.

1.1 Justificativa

O desenvolvimento de sistemas computacionais (*softwares*) é um campo que tem sido estudado com grande relevância na Ciência da Informação (CI). Um dos enfoques desse campo de estudo é a organização e uso da informação, no qual o protótipo[2] MHTX[3] merece destaque. O protótipo MHTX é um trabalho sobre a organização semântica do conhecimento, desenvolvido em uma pesquisa de doutorado pela professora Gercina Lima, em 2004, com a proposta de desenvolver um modelo de organização e recuperação de documentos baseado em atribuição semântica, através de *hiperlinks*.

Esse e outros tipos de sistemas de recuperação da informação evidenciam a tendência tecnológica para uma gestão informacional que se preocupe com a múltipla representação de significados (semântica) e com a sua recuperação através de interfaces mais amigáveis e intuitivas. Essa é uma realidade crescente nos Sistemas de Recuperação da Informação (SRI). Dessa forma, a necessidade de mecanismos de gerência e recuperação da informação em um contexto específico se faz mais presente a cada dia, principalmente a partir do crescimento informacional do período posterior à Segunda Grande Guerra. Foi nesse momento que estudiosos, como Bush (1945), vislumbraram modelos computacionais que poderiam contribuir para a complexa tarefa de controlar um acervo crescente de informações. Importantes teorias foram publicadas no sentido de

[2] Protótipo – instrumento que possui funcionalidades limitadas a fim de avaliar um projeto ou teoria.
[3] Mapa Hipertextual – um modelo para organização hipertextual de documentos.

aprimorar/aperfeiçoar a automação de processos de gestão informacional, como taxonomias e ontologias. Estes processos podem ser abordados com maior ênfase nos campos da linguística e da semântica.

Essa automação pode ser exemplificada através do modelo MHTX de recuperação semântica. Tal modelo avançou em diversos pontos, porém, no momento da presente pesquisa, encontra dificuldades em sua aplicação prática. Sua estrutura está implementada através de mais de um *software*, a fim de realizar processos informacionais dependentes na construção do acervo indexado. A estrutura de *softwares* descentralizados e independentes traz alguns inconvenientes para o profissional da informação responsável pela organização e indexação do acervo/da informação, visto que a informação é trabalhada em diferentes ambientes, cada um com suas particularidades. Organizar esse fluxo informacional torna-se, então, um desafio ainda maior, que poderia ser minimizado com o uso de uma estrutura única de *software*, em que os processos seriam trabalhados em um evento sequencial e dinâmico. O desenvolvimento desta estrutura constitui parte da proposta deste trabalho.

Atualmente, percebe-se que a recuperação da informação nos sistemas de recuperação é, muitas vezes, reduzida a palavras-chaves, que respondem a operadores booleanos[4] ou através de sistemas de aproximação sintática dos termos. Ao mesmo tempo em que a extensão informacional do documento permite um número maior de pontos de acesso, ela aumenta, igualmente, a complexidade para lidar com tal conjunto de informações. Dentro dessa realidade, tanto a capacidade do indivíduo de discernir a melhor estratégia de busca, quanto a quantidade de elementos a serem identificados como descritores resultam em um fator complicador no processo de indexação dos documentos. Observando-se as experiências da vida profissional e acadêmica do pesquisador, percebe-se que muitas informações não são encontradas devido à escassez de sistemas que promovam, com maior precisão, a interação entre usuário e acervo. Outro problema ocorre, com frequência, é a dificuldade de se realizar um processo indexador adequado à realidade informacional dos documentos em questão.

A grande quantidade de produção documental é outro fator que merece atenção. Tal volume impacta diretamente sobre o processo indexador, dificultando sua realização. Com a atual disponibilidade de mecanismos eletrônicos de baixo custo, os documentos são produzidos em uma velocidade que dificulta o trabalho de indexá-los à estrutura física e eletrônica disponível.

Diante do exposto, acredita-se que a proposta de um estudo direcionado aos metadados, atrelado à utilização de linguagens documentárias e estratégias de indexação

[4] Modelo de decisão que compreende apenas dois estados: verdadeiro ou falso.

semiautomatizada, pode contribuir sobremaneira em todo o processo de gerência de um acervo digital.

Considerando um acervo composto de processos como: escolha do item documental (seleção), inserção do item no acervo (indexação), armazenamento e posterior recuperação do item (pesquisa); a etapa da indexação deve receber especial atenção, pois ela influencia diretamente a recuperação da informação, seja feita através de linguagem natural ou através de um vocabulário. Uma indexação satisfatória inclui, entre outros fatores, a adoção de descritores adequados e o equilíbrio da quantidade desses descritores. A criação de mecanismos que facilitem e auxiliem o profissional responsável na realização do trabalho de indexação deve ser considerada como iniciativa relevante na Ciência da Informação.

Tendo em vista a necessidade do desenvolvimento de mecanismos facilitadores do trabalho do indexador, esta pesquisa pretende ampliar o entendimento sobre os metadados e sua aplicação em bibliotecas digitais, através de padrões e tecnologias que possam vir a implementar conceitos de gerência e recuperação de informação, em acervos digitais. Objetiva-se, ainda, a construção de um sistema de recuperação de informação, cuja base seja um mecanismo de linguagem documentária (taxonomia) que funcione como fator norteador do processo indexatório e que permita ao usuário definir os melhores termos e metadados que propiciem alcançar uma recuperação mais fácil e precisa dos documentos que respondem a determinada requisição.

1.2 Problema da pesquisa

Diante do aumento das publicações eletrônicas, do grande volume informacional e da crescente necessidade de se obter mecanismos de gerenciamento eletrônico para documentos em formato digital, os metadados têm um papel de destaque na gestão de acervos eletrônicos e manuais.

Pretende-se, ao final deste trabalho, responder à seguinte pergunta: Como os metadados podem ser utilizados na gestão de uma biblioteca digital e impactar na recuperação de itens específicos no acervo?

1.3 Objetivos

Este trabalho visa a pesquisar o potencial dos metadados no processo de gestão de arquivos em uma biblioteca digital e verificar sua aplicabilidade através da construção de um *software* inspirado no modelo MHTX de organização e busca semântica hipertextual de documentos.

1.3.1 Objetivo geral

Pretende-se estudar o potencial organizacional dos metadados em uma biblioteca digital e verificar suas funcionalidades através de um protótipo de *software* que realize o processo de gerenciamento e recuperação de documentos científicos.

1.3.2 Objetivos específicos

Foram definidos os seguintes objetivos específicos:

a) proporcionar maior eficácia, usabilidade e acessibilidade na organização e recuperação da informação através do uso de metadados em bibliotecas digitais;
b) disponibilizar uma ferramenta útil através da implementação de um *software* baseado no protótipo MHTX para gestores de conteúdo, autores de documentos, profissionais da área de informação e usuários.

1.4 Estrutura do trabalho

Além dessas considerações introdutórias, o presente trabalho está estruturado da seguinte forma:

O capítulo 2 descreve a importância das Bibliotecas Digitais, da indexação e das linguagens naturais ou limitadas a vocabulários no processamento informacional. Explora os metadados, objeto principal desta pesquisa, e inclui padrões importantes a serem considerados. Apresenta as tecnologias e padrões abertos para persistência[5] dos dados, em que tais aspectos de gravação influenciam a maneira como o acervo é indexado e recuperado.

O Capítulo 3 apresenta a metodologia utilizada para a realização da pesquisa.

O Capítulo 4 demonstra o desenvolvimento do protótipo, suas dificuldades e avanços na construção do *software*.

O Capítulo 5 apresenta as conclusões e apontamentos para pesquisas futuras.

[5] Gravação permanente de informações.

2 FUNDAMENTAÇÃO TEÓRICO-METODOLÓGICA

Neste capítulo, inicialmente serão apresentados alguns trabalhos relacionados com os estudos sobre Metadados aplicados a Bibliotecas Digitais, e a seguir os fundamentos teórico-metodológicos que embasaram este trabalho, destacando-se as bibliotecas digitais, os processos de indexação e catalogação, metadados, padrões e esquemas de gravação.

2.1 Fundamentos

O objetivo da organização informacional reside principalmente na necessidade de recuperação da informação armazenada. Para realizar essa atividade com maior rapidez e precisão, métodos e estratégias têm sido pesquisados no âmbito da Ciência da Informação a fim de permitir a eficiência desejada.

As Bibliotecas Digitais, inseridas nesse contexto de organização e recuperação, se apresentam como uma evolução das Bibliotecas tradicionais, influenciadas pelo recurso digital que inaugurou um novo modo de entender e trabalhar a informação.

Muitos autores, no intuito de contribuir com essa área científica, direcionaram seus esforços para as temáticas relacionadas à organização, indexação e recuperação da informação. Um importante recurso informacional que objetiva aumentar o potencial descritivo dessa grande massa documental, principalmente das novas mídias, são os metadados.

Os metadados são considerados por muitos pesquisadores como estruturas que fundamentam e possibilitam mecanismos descritores de documentos e mídias eletrônicas. Lourenço (2007, p. 72) destaca enfaticamente essas estruturas e considera sua presença essencial na sociedade da informação. A análise da autora destaca, ainda, a presença marcante no ambiente digital, sobretudo na *web*, de suas tipificações e padrões. Liu (2007, p. 132) expõe suas considerações sobre os metadados, enfatizando tipos, padrões e estruturas de codificação. Nos apontamentos direcionados à implementação e suas práticas, seus exemplos são voltados para Bibliotecas Digitais. Granitzer; Lux e Spaniol (2008, p. 4) escrevem uma obra composta por uma coletânea de capítulos de diversos autores, que abordam temas fortemente ligados às questões técnicas para conteúdo multimídia, seus padrões (MPEG) e enfoque semântico para organização e recuperação. Sicilia e Lytras (2009, p. 143) compilaram capítulos escritos por autores diversos e dividem seu livro em sete grandes blocos, abordando Herança cultural e preservação, Padrões de metadados e esquemas, Integração, *Web* semântica, Ontologias, Metadados em agricultura, e Semântica.

Rosetto (2003, p. 7) também realiza uma reflexão sobre o papel mediador que a informação em caráter descritivo possui, permitindo que os metadados forneçam uma representação da fonte, sua identificação e localização. Segundo a autora, os padrões de metadados possuem igual importância, pois a adoção de formatos é uma herança da história informacional. Marc, Foulonneau e Riley (2008, p. 117) destacam uma visão mais dissertativa, sobre criação, especialização e interoperabilidade, a ser disponibilizada pelos metadados.

O destaque sobre a necessidade e a importância da compatibilização de formatos de metadados também é uma preocupação compartilhada por Maurer e Nickerson (2007, p. 47), que observam a necessidade de se estudar a adoção de critérios práticos na transição de elementos de metadados de um padrão para outro. Alves e Souza (2007, p. 27), por sua vez, realizam um estudo objetivando a demonstração de correspondência entre dois padrões de extrema importância no campo da Ciência da Informação: *Dublin Core* e *Marc 21*. O documento das autoras remete à necessidade de ações para a implementação da interoperabilidade entre padrões de metadados, levando-se em consideração a necessidade de qualidade dos dados gerados a partir deste consenso/compatibilização e seus aspectos de precisão, fidelidade, seleção, generalizações, consistência, definições e fontes de dados. Sayão e Marcondes (2008) também destacam a necessidade de interoperabilidade nos ambiente de Bibliotecas Digitais e descrevem diversas tecnologias que podem contribuir para esse direcionamento.

A abordagem dos metadados, nesta dissertação, remete ao contexto das bibliotecas digitais. Dias (2006, p. 62) realiza uma importante descrição cronológica da evolução da biblioteca tradicional para a digital. Enfatiza, ainda, a importância do tratamento da informação e sua função descritiva, tanto do ponto de vista físico, como do ponto de vista temático, também chama a atenção para as dificuldades e desafios no ambiente das Bibliotecas Digitais. Cunha (2008, p. 4) também realiza uma reflexão sobre o contexto das bibliotecas tradicionais e digitais, ressaltando as possibilidades da representação no ambiente digital que, dificilmente, seria alcançado pela biblioteca convencional, essas possibilidades imprimem aspectos relacionados ao acesso à informação e à necessidade de integrar fontes e materiais eletrônicos. Além disso, pode-se notar que no aspecto econômico, o contexto digital pode ser viabilizado com um esforço muito menor do que no ambiente convencional.

Moraes e Oliveira (2010, p. 73) direcionam a atenção dos leitores para as Bibliotecas Digitais de Teses e Dissertações, destacando duas iniciativas: a Biblioteca Digital Brasileira de Teses e Dissertações (BDTD), do Instituto Brasileiro de Ciência e Tecnologia (IBICT), e o Banco de Teses da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). O documento retrata, sobretudo, os entraves para o

levantamento de teses e dissertações e aponta para eventuais problemas relacionados à autoria que podem vir a ocorrer. Entre eles estão: o eventual plágio de seus trabalhos, a exposição da qualidade de seus trabalhos, a perda de originalidade diante das exigências de ineditismo e também a perda de originalidade em relação à publicação de seu trabalho em formato de livro. Alguns problemas também foram encontrados com relação à versão do documento recuperado, algumas publicações foram disponibilizadas sem que houvesse uma revisão posterior à correção da banca.

Cunha (1999, p. 2) realiza uma investigação importante sobre a construção de uma biblioteca digital, elencando diversos pontos históricos e suas consequências no desenvolvimento no ambiente eletrônico. A abordagem apresenta, ainda, uma importante reflexão sobre a preservação da informação, um ponto muitas vezes ignorado pela facilidade em promover mecanismos de gravação em grande escala, visto que formatos eletrônicos e mídias estão mudando com uma velocidade que, no futuro, pode impedir a recuperação das informações. Um exemplo prático é o antigo disco flexível – disquete – que foi praticamente extinto do processo eletrônico atual.

O professor Soergel (2008, p. 6), ao se referir às características de uma biblioteca digital, suscita uma discussão, apresenta a sua concepção e segundo ele, aponta o mais importante: o questionamento sobre quais os componentes de sistema que, combinados, melhor suportam as necessidades do usuário e seu trabalho. Nessa linha de desenvolvimento, o autor chama a atenção para a organização do conhecimento e seus desafios no contexto das bibliotecas digitais.

A discussão sobre o processo de indexação é apresentada por diversos autores que se posicionam no sentido de contribuir com esta área, que é de suma importância no processo organizacional, considerando-se os contextos das bibliotecas tradicionais e digitais. Lancaster (2004, p. 24) possui uma obra que é referência no assunto na qual aborda com detalhes a teoria e prática da indexação. Lima e Boccato (2009, p. 136) realizam uma investigação importante sobre o processo de indexação e a geração de descritores, tendo como referência o Sistema Integrado de Bibliotecas da Universidade de São Paulo. O trabalho permitiu, entre outros aspectos, a comparação entre os mecanismos de indexação semiautomática e automática. Naves (2001, p. 193) apresenta um questionamento sobre o papel do indexador, suas práticas, experiências e limitações. O artigo convida o leitor a refletir sobre diferença do resultado no processo de indexação, ao ser submetido por indexadores com perfis profissionais diferentes. Gonçalves (2009, p. 97) também destaca a operação de representação documentária – indexação – em seu aspecto humano, ao colocar em evidência a percepção do usuário. Fujita (2009, p. 15) realiza uma investigação abrangente, que permite ao leitor refletir sobre diversos aspectos inerentes ao processo de indexação: especificidade, exaustividade, revocação e precisão. Além de

permitir uma análise mais profunda das diferentes perspectivas teóricas e metodológicas sobre esse processo.

O processo da presente pesquisa foi favorecido pela facilidade de consultas e recuperação de artigos e periódicos pela Internet, além da leitura de livros que abordam a temática da pesquisa, como os que foram apontados anteriormente.

Destaca-se a pesquisa de artigos na base do Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT) em sua Biblioteca Digital, enfatizando a própria biblioteca como instrumento de avaliação para o desenvolvimento deste trabalho.

A tese de doutorado da Profa. Dra. Gercina Lima (2004), que foi uma referência que fundamentou a concepção deste trabalho, tem como proposta o gerenciamento de conteúdo semântico com navegação em contexto, com a meta de otimizar os processos de organização, acesso e recuperação da informação em bibliotecas digitais, em qualquer área do conhecimento, possibilitando o gerenciamento do acervo por autores, gestores de conteúdo e profissionais da área da informação.

O protótipo é um modelo baseado no estudo ali dirigido e influenciou grande parte desta pesquisa.

Cabe valorizar ainda os artigos de Oddone e Gomes (2004), cuja Taxonomia, baseada no trabalho de Hawkins, foi utilizada nesta dissertação.

2.2 Bibliotecas digitais

Na trajetória da evolução humana, a importância da informação e sua preservação foram preocupações constantes. É notável como povos do passado remetem a essa perspectiva, através de sua constante busca por estabelecer uma escrita que pudesse ser codificada e decodificada nas eras vindouras. Vê-se, também, a evolução de seus instrumentos, para tornar essa informação permanente. Ao avançarmos da oralidade primária, a verbalização e seus desafios a fim de tornar a informação preservada entre as gerações, nos deparamos com o problema da “mídia” enquanto suporte: o mecanismo de armazenamento e de recuperação da informação.

Levy aponta essa oralidade, ao referir-se ao suporte enquanto humano, efetivado através de sua memória, o que torna possível o processo informacional: produção, codificação, armazenamento, decodificação, conforme se lê em:

> Numa sociedade oral primária, quase todo o edifício cultural está fundado sobre as lembranças dos indivíduos. A inteligência, nestas sociedades, encontra-se muitas vezes identificadas com a memória, sobretudo com a auditiva. (LEVY, 1993, p. 77).

Assim, verifica-se que, no processo de evolução, os mecanismos externos para armazenamento da informação, ou suporte (pedra, papiro, papel, fitas magnéticas, disquetes), e a própria linguagem se transformaram:

a) pictogramas e desenhos em cavernas, nas sociedades conhecidas como primitivas;
b) sistema de sinais criados pelos sumérios, armazenados em placas de argila;
c) papiro para armazenar os hieróglifos, além das esculturas e gravações em rochas e paredes nos monumentos egípcios;
d) a escrita Chinesa que, através de pictogramas, era gravada em cascos de tartarugas; mais tarde, é atribuída a esse povo a invenção do papel;
e) sistema de codificação/decodificação em fios com nós, denominado Quipos, usado pelos Incas;
f) o alfabeto, que revolucionou a história da informação.

Gutenberg desenvolveu a tipografia no período da Idade Média, criando o que podemos chamar de *industrialização da escrita*. Naquela época, livros, artigos, panfletos eram copiados manualmente por um grupo seleto de pessoas que detinham o conhecimento da codificação e decodificação da linguagem.

Burke, outro estudioso do assunto, opõe-se à imagem positivista da explosão informacional de Gutenberg. Segundo Burke,

> A imprensa foi descrita pelo humanista francês Guillaume Fichet – que introduziu a máquina impressora em Paris – como o “cavalo de Tróia”. Diferentes grupos sociais levantaram diferentes críticas ao novo instrumento. Por exemplo, os copistas e os “papeleiros” (que vendiam livros manuscritos) e os cantores contadores de histórias profissionais, todos temiam – como acontecera com os operadores de teares manuais na Revolução Industrial – que a imprensa os privaria de seu meio de vida. Os eclesiásticos, por sua vez, temiam que a imprensa estimulasse leigos comuns a estudar textos religiosos por conta própria em vez de acatar o que lhes dissessem as autoridades. (BURKE, 2002, p. 174).

Embora seja razoável considerar que a crescente publicação informacional possa criar alguns problemas, como, por exemplo, a dificuldade em gerir e encontrar a informação, sua forma em faceta eletrônica é exponencial, e, na sociedade atual, inevitável.

Percebendo esse processo, Vannevar Bush foi considerado o pioneiro do mecanismo organizacional, ao propor, em julho de 1945, uma máquina que pudesse organizar a informação da mesma forma que os humanos, o fazem: através de associações. Esse dispositivo foi denominado *Memex* e, de certa forma, assim como Charles Babbage e

sua máquina analítica, seu inventor não foi capaz de desenvolvê-la em sua plena capacidade de automação. Sua preocupação com a explosão informacional, ainda no pós-guerra de 1945, é determinante em seu trabalho:

> A ciência tem fornecido a mais rápida comunicação entre indivíduos, tem fornecido um registro de idéias e tem permitido ao homem manipular e fazer extratos daquele registro de modo que o conhecimento evolui e permanece durante toda a vida de uma raça e não a de um indivíduo.
>
> Há uma montanha crescente de pesquisa. Mas também há um aumento da evidência de que estamos atolando hoje à medida que a especialização se estende. O investigador é escandalizado pelos resultados e conclusões de milhares de outros trabalhadores, conclusões que ele não consegue encontrar tempo para entender, muito menos de lembrar como elas aparecem. No entanto, a especialização torna-se cada vez mais necessária para o progresso e os esforços para colmatar entre as disciplinas é, correspondentemente, superficial. (BUSH, 1945).

O computador, o novo veículo informacional, a faceta eletrônica, criada a partir da evolução digital, assemelha-se à revolução de Gutenberg[6], na qual a massificação da informação era vislumbrada através de um aparato tecnológico. A tecnologia intitulada digital permite transformar o mundo real em um conjunto de símbolos "virtuais" (intocáveis), codificados através de impulsos elétricos e simbolizados pelos sinais 0 (zero) e 1 (um), permitindo uma nova perspectiva na elaboração de acervos bibliográficos. Atualmente, esse conjunto não está mais preso fisicamente a um ambiente restrito, mas virtualmente expandido por um sistema interligado de computadores, permitindo seu acesso, de qualquer lugar, a qualquer pessoa que possua um dispositivo (computador, *palms*, *readerbooks*, entre outros) capaz de armazenar arquivos eletrônicos e apresentar, de forma gráfica/textual, o item pesquisado.

Todas essas possibilidades trazem consigo novos desafios para a Ciência da Informação, pois, com a facilidade de inclusão (cadastro) de novos itens no acervo digital, os processos de indexação e categorização precisam ser reavaliados, bem como o processo de recuperação. Ou seja, mesmo no ambiente digital, a correta abordagem dos processos de classificação e indexação é imprescindível para a sua posterior recuperação.

Totalmente incluídas nesse contexto, as Bibliotecas Digitais apresentam-se como resposta a uma demanda eletrônica que se acha em pleno crescimento no momento atual da história. Em países denominados "primeiro mundo", a versão eletrônica das publicações literárias já é disponibilizada como opção à versão impressa, e um novo negócio pode ser evidenciado com a crescente opção de livrarias virtuais pela Internet. Atrelado a esse momento propício à publicação e utilização do "eletrônico/virtual",

[6] Johannes Gensfleischzur Laden zum Gutenberg, de origem alemã, foi o inventor da impressão moderna baseada em tipos móveis.

dispositivos eletrônicos denominados *tablets* estão sendo lançados para prover uma condição simples e prática na leitura de informação digital e permitir práticas de interação com a Internet.

A fim de aprofundar a apresentação desse tema, apresentam-se, a seguir, algumas definições e características de Bibliotecas Digitais.

Dias possui a seguinte visão a respeito da biblioteca digital:

> Mas a biblioteca digital parece estar se firmando como a expressão que significaria, no contexto digital, um conjunto de artefatos, conhecimento, práticas e uma comunidade, que engendra compromissos realísticos assumidos por profissionais da informação, analistas de sistemas e usuários. (DIAS, 2001, p.á).

Para Marcondes, o conceito de biblioteca digital é mais próximo da tecnologia e se apresenta como:

> Biblioteca que tem como base informacional conteúdos em texto completo em formatos digitais – livros, periódicos, teses, imagens, vídeos e outros que estão armazenados e disponíveis para acesso, segundo processos padronizados, em servidores próprios ou distribuídos e acessados via rede de computadores em outras bibliotecas ou redes de bibliotecas da mesma natureza. (MARCONDES et al., 2006, p. 16).

Cunha caracteriza a Biblioteca Digital da seguinte maneira:

> A biblioteca digital combina a estrutura e a coleta da informação, tradicionalmente usada por bibliotecas e arquivos, com o uso da representação digital tornada possível pela informática. A informação digital pode ser rapidamente acessada em todo o mundo, copiada para preservação, armazenada e recuperada rapidamente. À semelhança da biblioteca convencional, a biblioteca digital também inclui os princípios consagrados de como a informação é organizada. (CUNHA, 2008, p. 5).

Sayão e Marcondes levantam a importância de a informação, na Biblioteca Digital, ser processada por diferentes *softwares* e seus fabricantes, fazendo referência à interoperabilidade:

> Idealmente, uma biblioteca digital deve ser capaz de armazenar uma variedade de tipos tradicionais de conteúdo – livros, periódicos, relatórios técnicos, *softwares* - bem como entidades multimídia complexas que misturam texto, imagens, vídeo e dados. Para que o acesso a essas informações seja efetivamente viável, o sistema no qual elas estão armazenadas deve ser capaz de gerar processos que sejam interoperáveis com os sistemas que estão à sua volta. (SAYÃO; MARCONDES, 2008, p. 136).

E Rosetto conceitua as bibliotecas digitais através da seguinte reflexão:

> A concepção de uma biblioteca digital deve ser realizada como uma ferramenta para propiciar o acesso à informação constituída em meio digital e também incluir outros meios tradicionais, mas, antes de tudo, deve constituir-se como um instrumento para a democratização do acesso ao conhecimento e inclusão social e cultural. (ROSETTO, 2008, p. 104).

Essa mesma autora também enfatiza a visão de que os metadados recebem um papel de destaque na recuperação das informações nesse ambiente:

> A criação de uma biblioteca digital implica conhecer todos os processos de tecnologia de informação (*hardware*, *software*, armazenamento, protocolos, etc.), e da biblioteca (definição do modelo de metadados, padrões a serem adotados, nível de detalhamento da descrição, metodologias para recuperar a informação organizada, entre outros requisitos), destacando-se os metadados (dados sobre dados), que serão a chave fundamental para proporcionar uma recuperação eficiente, eficaz e fácil de informações/documentos úteis para o usuário. (ROSETTO, 2003, p. 7).

A autora afirma, ainda, que essas estruturas informacionais possuem algumas características que merecem atenção:

> - Biblioteca digital tem também uma face de biblioteca e inclui coleções tradicionais e digitais, fixadas pelos meios tradicionais, ou seja, documentos impressos;
> - Biblioteca digital também inclui materiais digitais que existem fora de seu ambiente físico e administrativo, ou seja, em outras bibliotecas digitais e *websites*;
> - Biblioteca digital poderá incluir todos os processos e serviços que fazem parte da estrutura de bibliotecas. Entretanto, tais processos tradicionais, que fazem parte da base da biblioteca digital, terão que ser revisados e ampliados para acomodar as diferenças entre os novos meios digitais e os meios tradicionais;
> - Biblioteca digital, idealmente, proverá uma visão coerente de toda informação contida numa biblioteca, não importando a sua forma e formato;
> - Biblioteca digital servirá a suas comunidades específicas, assim como as bibliotecas tradicionais fazem agora, mas essas comunidades podem estar dispersas através da rede ou ampliadas;
> - Biblioteca digital requer habilidades de bibliotecários e de analistas de sistemas para serem viabilizadas. (ROSETTO, 2008, p. 104).

A biblioteca digital traduz diversos avanços e vantagens, se comparada à estrutura de uma biblioteca tradicional, tais como:

a) a facilidade de acesso, visto que não possui limites físicos/geográficos;
b) o baixo custo de implementação e manutenção;
c) a adoção do formato eletrônico pela maioria de editoras e autores;

d) a possibilidade de atualização constante.

Essas são características importantes que incentivam a adoção de tal tipo de estrutura em detrimento das estruturas físicas tradicionais. Um ponto que se destaca nessa reflexão está contido na observação de Dias:

> [...] comparado com o contexto tradicional, o que o contexto digital significa é um meio de facilitar o acesso a coleções que já existiam há muito tempo, com variada dificuldade de acesso, mas cujas eventuais facilidades providenciadas não podem competir com as extraordinárias facilidades que a Internet e a *Web* podem propiciar. (DIAS, 2001, p. 66).

Através dessas afirmações por diversos autores, pode-se entender que a dinâmica de alcance do ambiente eletrônico torna-se evidente em seus aspectos de acessibilidade e, muitas vezes, facilidade, se comparado ao ambiente tradicional, físico, das bibliotecas. O impedimento de acesso físico ao local pode ser transposto de modo extremamente simples no ambiente eletrônico, através das grandes redes de comunicação. Em contraponto, Dias ressalta a necessidade de cautela em relação a essa visão permeada de otimismo e entusiasmo em relação ao ambiente digital. Ele mostra que se deve levar em consideração "a ausência da característica de elemento de integração social e intelectual de suas respectivas comunidades e organizações, de presença marcante na biblioteca tradicional." (DIAS, 2001, p. 67). Ou seja, o ambiente digital não promove a participação coletiva dos indivíduos que procuram a informação, e encontro, debate e interação de ideias ainda não são práticas comuns nos sistemas de Biblioteca Digital, na qual apenas a disponibilização do acervo para solicitações individuais é alcançada.

2.2.1 Biblioteca Digital de Teses e Dissertações - IBICT

No Brasil, um ótimo exemplo de Biblioteca Digital pode ser consultado no IBICT – Instituto Brasileiro de Informação em Ciência e Tecnologia. Seu enfoque é centrado na gestão de um acervo nacional de Teses e Dissertações. O modelo do IBICT é descrito por Sílvia Southwick:

> A BDTD[7] adota um modelo distribuído utilizando-se das tecnologias de arquivos abertos[8]. As instituições de ensino superior (IES) são provedores de dados [...] e o IBICT opera nessa rede como agregador, coletando metadados de teses e dissertações dos provedores de dados, provendo serviços de informação sobre esses metadados e expondo esses

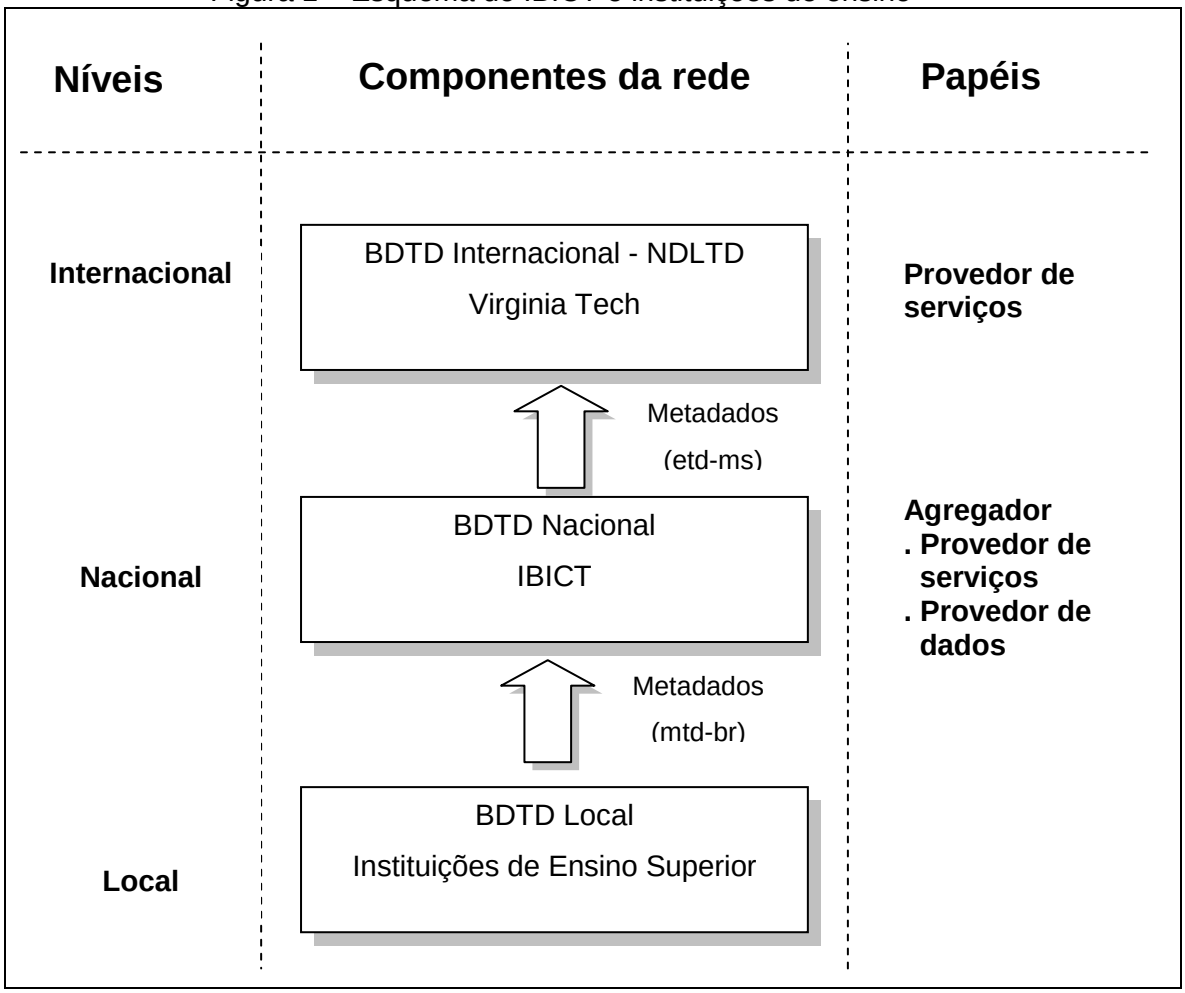
[7] Biblioteca Digital Brasileira de Teses e Dissertações (INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA E TECNOLOGIA, 2012).

[8] (*OPEN ARCHIVES*, 2012).

> metadados para serem coletados por outros provedores de serviços. Em especial a BDTD expõe metadados para serem coletados pelo provedor de serviços internacional NDLTD[9] (*Networked Digital Library of Thesis and Dissertation*). (SOUTHWICK , 2003, p. 3).

Esse modelo pode ser entendido graficamente através do seguinte esquema:

Figura 1 – Esquema do IBICT e instituições de ensino



Fonte: SOUTHWICK , 2003, p. 4

O esquema, mostrado na figura 1, evidencia que as instituições de ensino promovem a divulgação dos metadados e a exposição do documento (tese/dissertação) em seu servidor de arquivos. A BDTD do IBICT recebe os metadados e disponibiliza uma interface de busca/recuperação para os visitantes através de seu site na Internet. Em outra etapa, esse mesmo mecanismo disponibiliza os metadados acrescidos de outras informações para outra biblioteca, a *Virginia Tech*, que pretende agregar um acervo internacional de teses e dissertações.

[9] Organização internacional dedicada a promover a adoção, criação, utilização e preservação de Teses e Dissertações eletrônicas. (*NETWORKED DIGITAL LIBRARY OF THESES AND DISSERTATIONS*, 2012).

O modelo do IBICT, exemplificado neste capítulo e mostrado na Figura 1, permite observar a preocupação com diversos elementos importantes na criação de uma biblioteca digital. Pode-se perceber, portanto:

a) a definição de um padrão de metadados baseado em *Dublin Core* e expandido com mais elementos;
b) a adoção de um padrão aberto de arquivos entre as instituições;
c) a disponibilização, na Internet, de uma interface de fácil acesso independente de *software* que permita a interoperabilidade.

Além dessas considerações, o IBICT expandiu o padrão de metadados internacional etd-ms (baseado em *Dublin Core*), criando o seu padrão: mtd-br. A plataforma da solução foi desenvolvida no ambiente *web* (Internet), o que facilita a integração com as demais instituições de ensino, porém, a metodologia de catalogação, a indexação do acervo e o mecanismo de *ranking*, o qual permite ordenar a listagem de resposta, não são expressos na documentação.

O IBICT é uma referência nacional em Biblioteca Digital, mas cabe ainda identificar outras iniciativas como a da Biblioteca Digital da UFMG, Figura 2, ainda em período de implementação no momento desta pesquisa, que, inicialmente, disponibiliza, em seu acervo, teses e dissertações defendidas nas áreas de Ciência da Informação e Linguística.

Figura 2 - Biblioteca Digital da Universidade Federal de Minas Gerais

Fonte: UNIVERSIDADE FEDERAL DE MINAS GERAIS, 2012

Outras de bibliotecas digitais que merecem destaque:

a) Biblioteca Nacional Digital do Brasil (BIBLIOTECA NACIONAL, 2012) - em que podem ser encontrados fotografias e livros raros digitalizados, entre outros materiais de grande valor cultural;
b) Portal Domínio Público, iniciativa do Ministério da Educação (BRASIL, 2012a) - com destaque para o acervo completo de Machado de Assis, entre outros clássicos;
c) Biblioteca Digital de Teses e Dissertação da USP (UNIVERSIDADE DE SÃO PAULO, 2012) - que disponibiliza teses e dissertações defendidas na instituição;
d) Biblioteca Digital da Unicamp (UNIVERSIDADE DE CAMPINAS, 2012) - que também disponibiliza teses, dissertações e produções técnico-científicas digitais produzidas na instituição;
e) Biblioteca Digital do Senado Federal (BRASIL, Senado Federal 2012b) - que permite o acesso a mais de 200 mil documentos (no momento desta pesquisa) de interesse do Poder Legislativo Brasileiro.

Pode-se observar que as bibliotecas digitais brasileiras estão crescendo em número e em qualidade de interação com o usuário, a fim de disponibilizar seu acervo e, independente do modelo tecnológico adotado, é necessária uma contínua reflexão sobre o processo de indexação; processo essencial na organização em qualquer sistema de recuperação da informação.

2.3 Indexação

Esta seção pretende esclarecer aspectos importantes do processo de indexação, também conhecido como catalogação de assuntos, na Ciência da Informação (FUJITA, 2009, p. 14). Pretende-se apresentar, também, o impacto desta etapa no projeto em desenvolvimento, assim como as estratégias para utilizar linguagens documentárias, vocabulários controlados ou estruturas de organização como alternativa auxiliar na etapa do tratamento do item no acervo, bem como na recuperação da informação.

Os profissionais da ciência da informação buscam trabalhar com a organização e o acesso aos sistemas de informação, habilitando-se para a classificação, a indexação e a busca em sistemas informatizados (LIMA, 2009, p. 99) A indexação compreende a identificação do conteúdo do documento por meio do processo de análise de assunto e

também a representação desse conteúdo por meio de conceitos (RUBI, 2009, p. 81). Esse processo tem, como objetivo principal, fazer coincidir a linguagem do sistema com a linguagem do usuário final.

A indexação deve ser considerada um fator norteador em todo o processo informacional na gestão de um acervo, porque é a partir da organização inicial, fruto da indexação, que os processos de organização e recuperação localizam o item pesquisado. Segundo Borko,

> Sem índices, nós não poderíamos telefonar facilmente, fazer pedidos em restaurantes, programar uma viagem, encontrar um material dentro de uma biblioteca, agendar tarefas, localizar correspondências, procurar ruas, ou realizar um grande número de outras coisas que sabemos ter importância. (BORKO, 1978, p. 3).

Também em Lancaster, é perceptível a avaliação de Borko, na qual os demais processos dependem da indexação, pois:

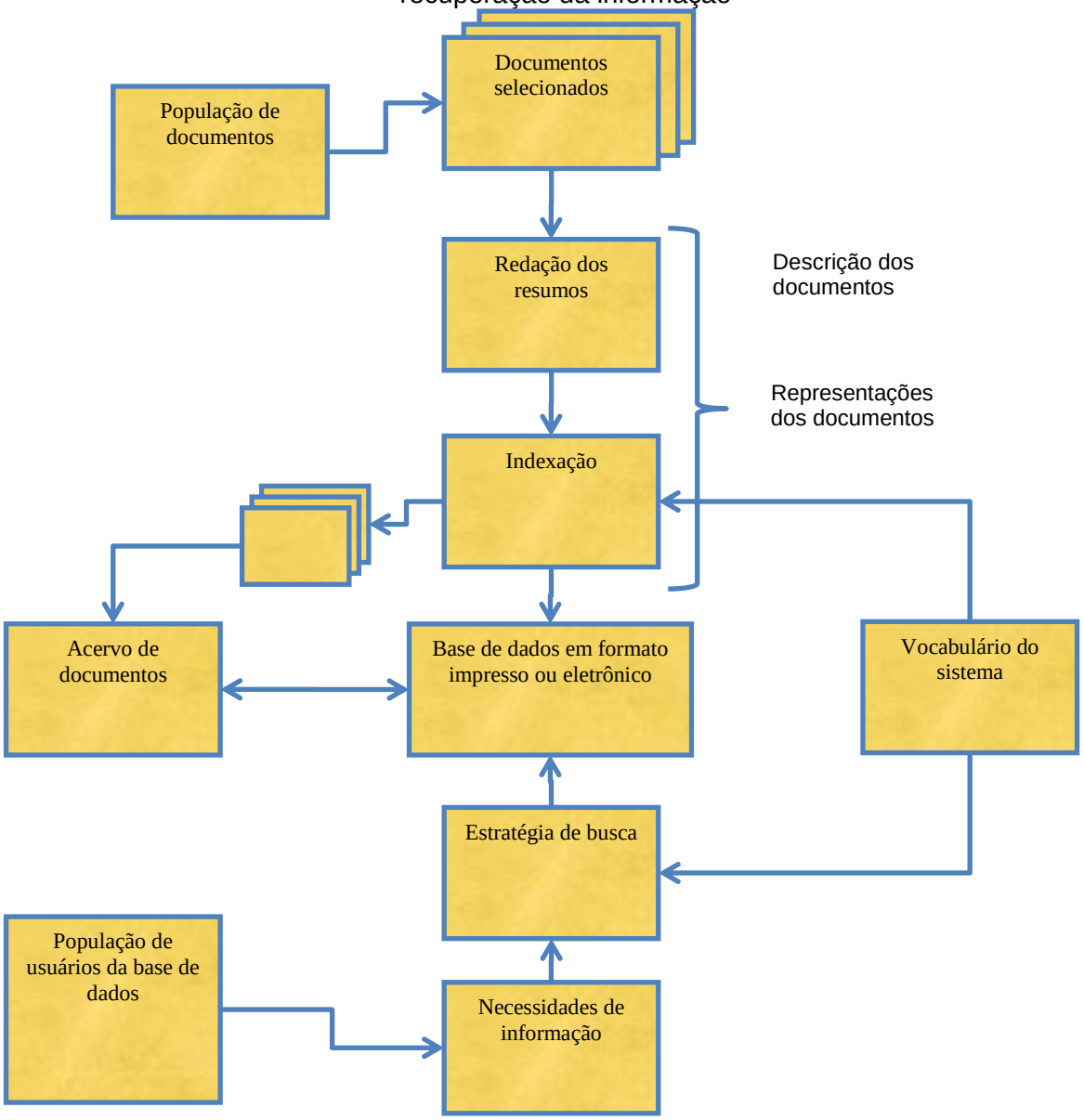
> Os termos atribuídos pelo indexador servem como pontos de acesso mediante os quais um item bibliográfico é localizado e recuperado, durante uma busca por assunto num índice publicado ou numa base de dados legível por computador. (LANCASTER, 2004, p. 24).

Robredo concorda com Borko e Lancaster ao considerar a influência da indexação:

> Indexação é um processo intelectual que pressupõe que o acesso à informação documentária, por intermédio dos termos – ou dos códigos – de indexação, será o ponto de partida para selecionar os próprios documentos. [...] A indexação consiste em indicar o conteúdo temático de uma unidade de informação, mediante a atribuição de um ou mais termos (ou códigos) ao documento, de forma a caracterizá-lo de forma unívoca. A finalidade do processo de indexação é a recuperação da informação para satisfazer as necessidades dos usuários potenciais. (ROBREDO, 2005, p. 165)

Em sua obra *Indexação e Resumos*, Lancaster elaborou um fluxograma do processo de indexação. A figura ilustra a entrada, o processamento e a saída da informação, enfatizando o procedimento da indexação como fator determinante da base de dados e sua consequente influência na escolha da estratégia de recuperação da informação.

Figura 3 - A função da elaboração de índices e resumos no quadro mais amplo da recuperação da informação



Fonte: Lancaster (2004)

A indexação feita por seres humanos é um processo intelectual subjetivo, no qual os indexadores nem sempre incluem um assunto que deveria ser incluído, representam um assunto com o melhor termo possível ou explicitam alguma relação de interesse potencial para certos usuários (LANCASTER, 2004, p. 256). Dessa forma, a prática de indexar deve ser valorizada, principalmente no momento em que a informação precisa ser recuperada do acervo em que se acha armazenada. Uma boa indexação permite uma recuperação mais simples e assertiva, enquanto um processo pouco adequado resulta em uma dificuldade maior em encontrar o item desejado.

Na indexação manual, o indexador, em um primeiro momento, realiza a leitura documentária para a identificação e a seleção dos conceitos expressos em um documento

e, a seguir, representa (“traduz”) esses conceitos selecionados em descritores da linguagem documentária adotada pelo Sistema de Informação. Para que uma indexação expresse de maneira completa o conteúdo de um documento, exige-se do indexador um conhecimento prévio da área de domínio e do próprio documento. Lancaster (2004, p.24) cita partes relevantes do texto a serem observadas nesse processo, uma vez que o tempo de que o indexador dispõe não lhe permite realizar uma leitura profunda e completa do texto a ser indexado.

São elementos de maior importância na leitura para se realizar a indexação:

a) título;
b) resumo, se houver;
c) sumário;
d) introdução, frases e parágrafos de abertura de capítulos, e as conclusões;
e) ilustrações, gráficos, tabelas e respectivas legendas;
f) palavras ou grupos de palavras que apareçam sublinhados ou grafados com tipos diferentes.

Quanto à dinâmica manual da indexação, Lancaster acrescenta que:

> Os indexadores humanos procurarão selecionar expressões do texto que pareçam ser bons indicadores daquilo de que trata um documento. Provavelmente serão influenciados pela frequência com que um termo aparece no documento e talvez onde apareça – no título, resumo do autor, legendas das ilustrações, etc. – e por seu contexto. (LANCASTER, 2004, página).

Na indexação semiautomática, é realizado um processo misto entre processamento computacional e decisão humana. O usuário pode marcar partes do texto/documento e comandar a máquina (*software*) para realizar a construção do índice, com base nos termos selecionados.

Outra técnica do modelo semiautomático consiste no levantamento do número de ocorrências de determinados termos – organizados em um vocabulário controlado – através de algoritmos, para detecção de repetição, com posterior realização de uma escolha manual para eleger os termos encontrados como melhores descritores.

Os desafios do processo semiautomático, assim como na indexação automática, figuram na dificuldade computacional em lidar com informações de significado semântico, polissemias e regionalismos. Cabe ao indexador a contextualização da informação, o que minimiza os problemas de automatismos completos.

Na indexação automática, a máquina realiza o processo de representação do texto com base em um sistema de regras pré-definidas através de algoritmos (comandos que constituem o *software*), pelos quais o documento é associado a palavras ou termos descritores. Uma técnica conhecida na realização desse processo é a construção de um índice invertido[10] a partir dos termos do próprio documento, ignorando os termos, sílabas e palavras indesejadas (*stop words*) na recuperação. O sucesso dessa abordagem depende de uma série de fatores, dentre eles: o vocabulário controlado utilizado para a submissão do documento, a identificação de elementos (área) importantes dentro do documento, como títulos, palavras-chaves, resumo, entre outros, ou mesmo a construção de índices mais completos.

Entende-se ainda que, em todos os modelos de indexação, a questão semântica é um problema a ser enfrentado, devido à particularidade de cada língua, que, em razão da sua dinâmica, determina constante modificação no significado das palavras. Nesse sentido, estudos vêm sendo realizados a fim de confrontar questões linguísticas com a realidade da Ciência da Informação, resultando em ferramentas como os tesauros[11] e as ontologias[12], que oferecem um referencial nos avanços para esses desafios.

O presente trabalho adotou como foco de estudo, a submissão de um dado documento a uma taxonomia, adotando um sistema de peso (importância) associado à posição no texto do termo encontrado, permitindo a inferência humana como fator decisório no processo de eleição dos melhores termos, caracterizando o processo como uma indexação semiautomática.

Percebe-se, em todas as técnicas e projeções estudadas, que o problema da subjetividade humana é inevitável e reside em todo o processo, seja na construção/definição da linguagem documentária, seja na definição dos parâmetros do documento a serem pesquisados, ou mesmo na definição das regras do algoritmo computacional a ser criado. Isso faz com que o mesmo documento possa ser (e provavelmente será) recuperado através de diferentes representações em diferentes modelos de indexação, seja ele manual, semiautomático ou automático.

É importante observar que, independentemente do modelo de indexação utilizado, todo o processo de organização da informação é baseado em linguagens, as quais expressam capacidade de comunicação e estabelecem um padrão entre autor e leitor, seja através de uma leitura textual, simbólica ou composta por várias mídias.

---

[10] Estrutura computacional na qual o termo aponta para a referência de sua localização em um ou vários documentos.

[11] Linguagem documentária dinâmica que contém termos relacionados semântica e logicamente, cobrindo de modo compreensivo um domínio do conhecimento. (GOMES, 1990, p.16)

[12] Modelo que permite explicitar as relações entre os dados em um determinado domínio de conhecimento.

2.4 Linguagens

Toda a elaboração do processo de indexação tem como objetivo criar uma representação da informação, a qual pode ser construída através de um mecanismo de padronização e organização, denominado vocabulário controlado. Essa estrutura é constituída por um conjunto de termos previamente autorizados em uma unidade de conhecimento, com vistas a minimizar os problemas relacionados a sinônimos, homônimos e expressões semânticas.

Tal vocabulário não deve ser visto apenas como uma lista de palavras/termos selecionados em um determinado contexto (universo de estudo), mas pode incluir no mesmo instrumento, recursos para tratamento de significado (semântica) e, dependendo do seu tipo permite:

> a) controlar sinônimos ao escolher uma única forma de representação daquele conceito;
> b) diferenciar homógrafos, palavras com a mesma sintaxe com sentido/significado totalmente diferente, exemplo Peru (país/ave);
> c) reunir ou ligar termos cujos significados apresentem uma relação de proximidade. Dois tipos de relações são identificados explicitamente: as hierárquicas e as não-hierárquicas (ou associativas). Por exemplo, o termo mulheres operárias relaciona-se hierarquicamente com mulheres (como uma espécie desse termo) e com donas de casa (também uma espécie do termo mulheres), bem como está associado a outros termos, como emprego ou famílias monoparentais, que aparecem em hierarquias bem diferentes. (LANCASTER, 2004, p. 14).

Percebe-se que a norma ANSI/NISO Z39.19-2005[13] possui concordância com os princípios descritos por Lancaster, uma vez que define como propósito dos vocabulários controlados:

> 1. Tradução: fornecer um meio para converter a linguagem natural dos autores, indexadores e usuários em um vocabulário que pode ser utilizado para indexação e recuperação;
> 2. Consistência: promover a uniformidade no formato e na atribuição de termos;
> 3. Indicação de relacionamentos: indicar relações semânticas entre os termos;
> 4. Etiqueta e navegação: fornecer hierarquias coerentes e claras em um sistema de navegação para ajudar os usuários a localizar os objetos de conteúdo desejado;
> 5. Recuperação: serve como um auxílio na localização de objetos em busca de conteúdo. (LANCASTER, 2004, página).

[13] Esta norma estabelece diretrizes para a construção, formato e gestão de vocabulários controlados monolíngues.

Lancaster enumera três tipos de vocabulários controlados:

> Esquemas de classificação bibliográfica (como a Classificação Decimal de Dewey)[14]; Listas de cabeçalhos de assuntos e Tesauros. (LANCASTER, 2004, p. 19).

Todos esses instrumentos procuram listar seus termos em formas alfabéticas e também sistematicamente.

Por sua vez, a norma Z39.19-2005 inclui o instrumento Tesauro também como um tipo de vocabulário controlado.

No esquema de classificação bibliográfica, a organização alfabética ocorre em segundo plano, em forma de estrutura que aponta para a organização principal, hierárquica, como na figura abaixo:

Figura 4 – Classificação Decimal de Dewey

	Dewey Decimal System
000	Computer science and information
100	Philosophy and psychology
200	Religion and mythology
300	Social sciences
400	Language
500	Science and math
600	Technology
700	Arts and recreation
800	Literature
900	History and geography

Fonte: <http://www.bridgeviewlibrary.org/bridgeview/adult_dewey.asp>

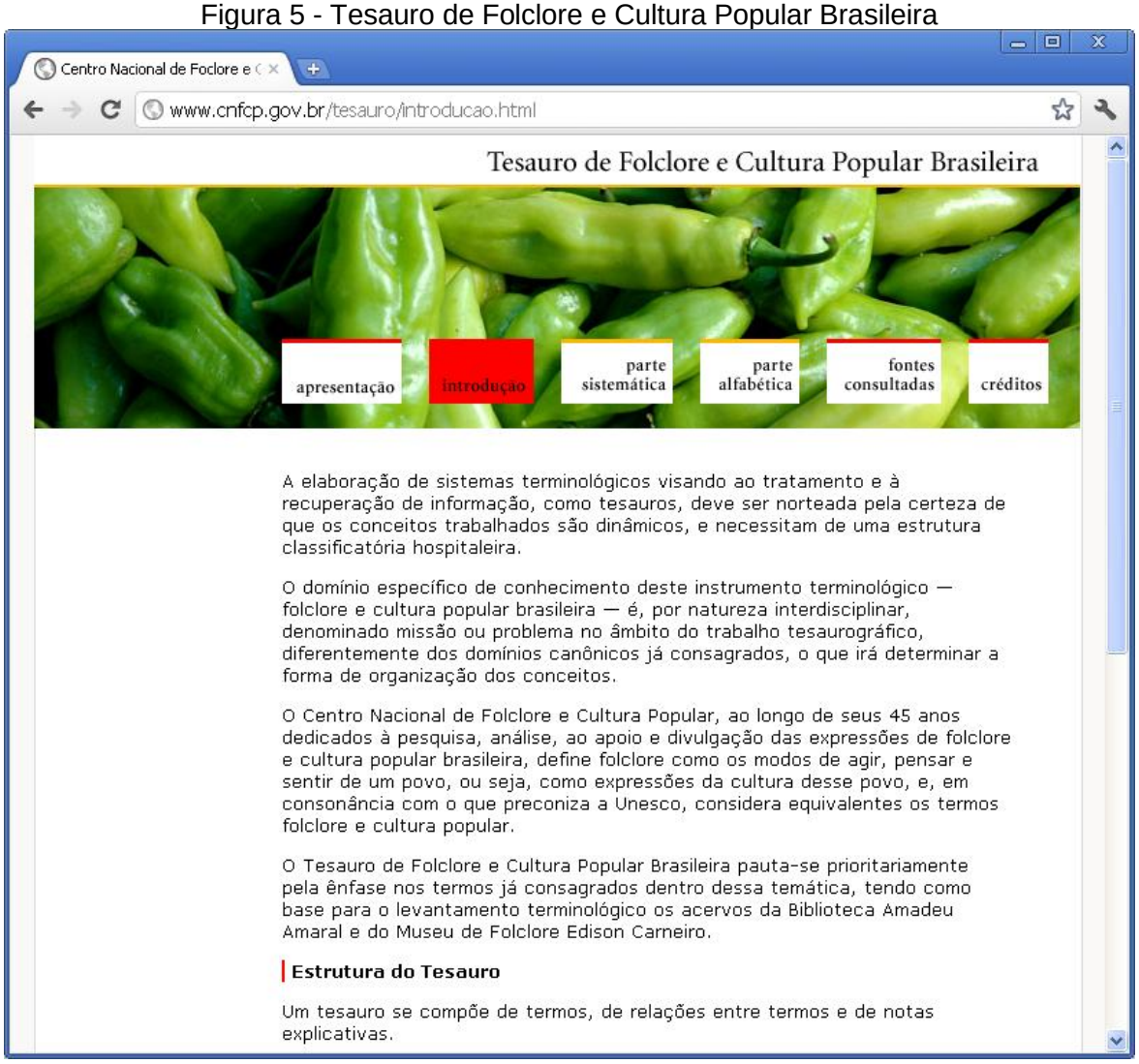
No Tesauro, Figura 5, a organização dos termos é alfabética, mas há uma estrutura hierárquica implícita que retoma referências remissivas. As siglas indicam, respectivamente:

BT – *Broader Term* (termo genérico) – Pode ser traduzido como termo descritor superior ao termo em questão na hierarquia do tesauro. Observa-se que termos descritores na condição de BT, também são chamados na literatura de termos de topo. Existem tesauros que aceitam relacionamento poli-hierárquico.

NT – *Narrower Term* (termo específico) – Como o próprio nome indica, é um termo específico e está sempre associado a um BT. Um NT pode ter outros NTs abaixo dele, para os quais se torna um BT.

[14] Sistema de classificação documentária amplamente utilizado, que organiza o conhecimento utilizando dez classes principais.

Figura 5 - Tesauro de Folclore e Cultura Popular Brasileira



Fonte: <http://www.cnfcp.gov.br/tesauro/alfabetica.html>

As listas de cabeçalhos de assuntos assemelham-se a um Tesauro por se basearem na ordenação alfabética, mas, segundo Lancaster (2004), elas implementam uma estrutura hierárquica imperfeita e não fazem distinção clara entre relações hierárquicas e associativas.

Os três instrumentos (vocabulário controlado, tesauro e lista de cabeçalho) possuem estratégias bem diferentes para implementar seus vocabulários. É importante ressaltar que um vocabulário controlado visa a melhorar o processo de indexação em relação a um grupo de documentos inserido em um domínio de conhecimento. Ou seja, quando pessoas diferentes realizam o processo de indexação, a possibilidade de cada indexador interpretar os termos livres de maneira diferente é muito grande. Mesmo utilizando um vocabulário controlado, essa diferença existe na escolha do melhor identificador, mas a presença do vocabulário minimiza a diferença, pois reduz o número de

opções e elimina os problemas de sinônimos, exercendo uma influência positiva no processo de indexação.

Pode-se entender, ainda, que, quanto mais pontos de acesso um determinado documento possuir, mais alta será a revocação, porém, menor sua precisão, pois o fato de o documento ser descrito através de um número maior de termos evidencia a inclusão de descrições secundárias. Segundo Lancaster (2004, p. 255), um maior número de pontos de acesso provê maior probabilidade de relações ilegítimas, a saber: falsas associações e relações incorretas entre termos.

Deve-se considerar a existência de outros instrumentos além dos citados por Lancaster, como as Taxonomias, instrumento escolhido neste trabalho para ser utilizado no protótipo.

2.5 Taxonomias

As Taxonomias são estruturas classificatórias que atuam na organização e recuperação da informação, porém sem o controle de sinônimos e homógrafos, realizado pelos instrumentos anteriores. Neste trabalho, as Taxonomias possuem grande importância em razão de sua aplicação prática no protótipo. Segundo Campos e Gomes,

> Recentemente, o uso de taxonomias tem sido adotado por permitir acesso através de uma navegação em que os termos se apresentam de forma lógica, ou seja, em classes, sub-classes, sub-sub-classes, e assim por diante, em quantos níveis de especificidade sejam necessários, cada um deles agregando informação sobre os documentos existentes na base. Uma vantagem desta forma de acesso é a garantia, para o usuário, da melhor seleção do termo de busca, uma vez que as classes contêm tópicos mutuamente exclusivos.
> [...]
> No âmbito da Ciência da Informação as taxonomias podem ser comparadas a estruturas classificatórias como as Tabelas de Classificação, que têm como objetivo reunir documentos de forma lógica e classificada. Atualmente, as taxonomias reúnem todo tipo de documento digital e permitem, diferentemente das estratégias de busca, um acesso imediato à informação. Ao contrário das Tabelas, que oferecem um endereço (notação) que localiza os documentos nas estantes, a taxonomia prescinde de notação. (CAMPOS; GOMES, 2008).

As taxonomias surgiram nas áreas de Botânica, Zoologia e Paleontologia, e são estruturas base para o trabalho moderno de diferentes ciências, naturais e sociais (VICKERY *apud* ODDONE; GOMES, 2004). Elas precisam estar contextualizadas e são restritas a uma área de conhecimento ou domínio.

No presente trabalho, o instrumento taxonômico será adotado como estrutura organizadora de termos que vão influenciar a indexação e, consequentemente, a recuperação da informação:

a) Entrada: o documento selecionado para ser adicionado deve ser do mesmo domínio do vocabulário controlado.
b) Processamento: para representar a informação gravada em documentos na biblioteca digital, termos da taxonomia devem ser associados ao documento a fim de criar a descrição e permitir a recuperação futura.
c) Saída: para recuperar o documento, o usuário poderá utilizar termos da Taxonomia, evitando problemas de sinonímia, sintaxe e semântica, por exemplo.

A adoção de uma Taxonomia visa a minimizar diversos problemas, entre eles, o da subjetividade. Embora esse instrumento não trate homônimos, a delimitação da área do conhecimento já atua como elemento limitador de problemas relacionados à semântica. Outro desafio a ser considerado quando do uso desse instrumento é a inclusão de novos termos no domínio do assunto, o que exige dinâmica na atualização e na manutenção dos seus elementos.

Existem situações em que a linguagem natural possibilitará uma melhor precisão, visto que a taxonomia sempre sujeita o processo de pesquisa aos termos pré-definidos. Certamente, essa situação depende da política de indexação adotada pelo sistema de recuperação da informação. Lancaster (2004, p. 257) afirma ainda que o uso de vocabulário controlado é, muitas vezes, a escolha dos especialistas em informação que conhecem bem as diretrizes e regras desse instrumento, enquanto a linguagem natural tem a preferência de usuários em um determinado assunto.

Neste trabalho, a proposta da adoção do instrumento taxonômico permitirá a atualização e a personalização do instrumento, bem como a definição de novas Taxonomias pelo usuário. A linguagem natural será abordada como segunda opção de recuperação, o que permitirá a recuperação em metadados não abordados pela Taxonomia, objetivando aspectos anteriormente citados.

Objetivando maior possibilidade de recuperação de informações, os termos do vocabulário serão confrontados com o acervo, realizando-se uma pesquisa bem mais assertiva em comparação com a linguagem natural, na qual os termos e sua construção dependem de vários fatores relacionados ao usuário, como o conhecimento prévio do assunto, a utilização de sinônimos e a combinação de termos.

Uma Taxonomia importante sobre a Ciência da Informação foi criada por Oddone e Gomes (2005) e será utilizada como instrumento inicial nas configurações do protótipo. Ela se encontra como anexo (Anexo A) deste trabalho.

2.6 Metadados, padrões e linguagem

Uma questão importante a se pensar é o fato de que as diferentes manifestações das mídias podem não possuir um conteúdo textual, como é o caso de arquivos de áudio, vídeo e imagens. Apesar disso, no entanto, é preciso que a escrita, sistema utilizado pelas linguagens de recuperação, seja inserida de alguma maneira no item a ser recuperado. E, mesmo quando o item possui um conteúdo textual, a terminologia utilizada para a sua recuperação pode ser aprimorada pelos metadados.

2.6.1 O que são os metadados

Para realizar a gestão de um acervo, seja ele físico ou digital, são necessárias estruturas com a função de representar/descrever os itens que compõem essa coleção. A estrutura mais utilizada com esse propósito é denominada “metadados”, que significa “dados sobre dados”, ou seja, informação que descreve outra informação (JIALIU, 2007).

Rosetto (2003) estabelece um quadro descritivo em que conceito, objetivos e características dos metadados são mais bem evidenciados:

Quadro 1 - Conceito, objetivos e características de metadados

Conceito	Metadados são um conjunto de dados-atributos, devidamente estruturados e codificados, com base em padrões internacionais, para representar informações de um recurso informacional em meio digital ou não digital, contendo uma série de características e objetivos.
Objetivos	Localizar, identificar e recuperar dados de um recurso informacional. Propiciar controles de ordem gerencial e administrativo permitindo conexões e remissões (*links*) para pontos internos e externos ao sistema. Possibilitar a interoperabilidade entre sistemas de informação, dentro de padrões. Informar sobre as condições de acesso e uso da informação. Ser legível tanto pelo homem como pela máquina. Possibilitar a elaboração de índices.
Características	Permitem a descrição, com pormenores, das condições físicas dos componentes com o fim de identificar e caracterizar o recurso de informação. Observância de padrões internacionais para a sintaxe e semântica da especificação do recurso de informação, em meio digital ou não digital. Informam sobre a armazenagem, preservação, acesso e uso dos dados. Dispõem informações administrativas e gerenciais para a devida criação e definição de responsabilidades dos metadados.

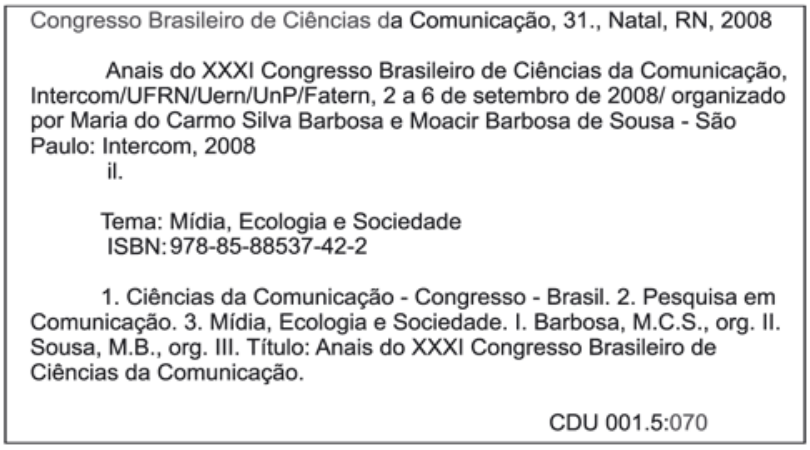
	Possibilitam análises da qualidade, avaliações e formas de uso. Auto-descrevem e criam documentação própria que subsidia o gerenciamento dos recursos informacionais.

Fonte: Rosetto (2003)

Ao abordar os metadados na perspectiva do controle manual de um determinado acervo bibliográfico, as fichas catalográficas são um exemplo bastante claro sobre o funcionamento desse tipo de controle. Ainda hoje, em algumas bibliotecas, para organizar e identificar um recurso bibliográfico utiliza-se a ficha catalográfica. Esse modelo permite obter um conjunto abrangente de informações sobre a publicação, possibilitando que o profissional da informação realize o processo de catalogação e recuperação do item dentro de um determinado acervo. Tradicionalmente, a ficha catalográfica (figura 6) é disponibilizada em formato físico, com informações impressas em formato retangular. Ela representa um item do acervo, através de um conjunto de elementos que pretendem identificá-lo de forma única.

Pode-se entender, portanto, que a ficha ou registro catalográfico é uma estrutura ordenada e bem definida para a organização da informação. Esses elementos de dados – objetos da catalogação – oferecem, ao profissional catalogador, a possibilidade de organizar a informação de determinado recurso bibliográfico de modo regular, sendo que a padronização universal de tal conjunto de elementos é um desejo antigo. Muitos acervos possuíam certo nível de organização descritiva, porém, sem a possibilidade de intercâmbios efetivos, visto que cada acervo contava com uma lógica descritiva e um número de elementos que não eram compatíveis entre as bibliotecas.

Figura 6 – Exemplo de ficha catalográfica


Congresso Brasileiro de Ciências da Comunicação, 31., Natal, RN, 2008

Anais do XXXI Congresso Brasileiro de Ciências da Comunicação, Intercom/UFRN/Uern/UnP/Fatern, 2 a 6 de setembro de 2008/ organizado por Maria do Carmo Silva Barbosa e Moacir Barbosa de Sousa - São Paulo: Intercom, 2008
il.

Tema: Mídia, Ecologia e Sociedade
ISBN: 978-85-88537-42-2

1. Ciências da Comunicação - Congresso - Brasil. 2. Pesquisa em Comunicação. 3. Mídia, Ecologia e Sociedade. I. Barbosa, M.C.S., org. II. Sousa, M.B., org. III. Título: Anais do XXXI Congresso Brasileiro de Ciências da Comunicação.

CDU 001.5:070

Fonte: Congresso Brasileiro de Ciências da Comunicação, 2008

Todo este processo manual foi alterado com a evolução da computação, a ficha catalográfica foi convertida em formato legível, passível de processamento para o

computador, ou seja: no processo digital advindo do período histórico da computação, uma nova abordagem se tornou possível, uma vez que a capacidade de memória e processamento passou a contar com o auxílio de uma máquina capaz de processar e armazenar um conjunto de informações gigantesco.

Nesse aspecto, em um primeiro momento, a informática funcionou como uma simples máquina de escrever com capacidade de armazenamento. No entanto, cientistas da computação e da informação vislumbraram um potencial infinitamente maior, pois essa máquina podia realizar inferências e até tomar decisões. Naturalmente, acervos começaram a ser geridos com o auxílio da computação, e a ficha catalográfica era inserida em programas conhecidos como banco de dados, que são estruturas organizacionais baseadas em um modelo de entidades[15] e relacionamentos[16]. Nesse universo, o poder de recuperação tornou-se evidente. Cabe ressaltar que o aspecto técnico-computacional ainda causa um entusiasmo latente no leitor, no usuário, naquele que manuseia uma máquina que, a princípio, "tudo responde, tudo sabe". Mas, na realidade que antecedeu a digitalização, muitas vezes o processo intelectual no discernimento da informação, da criação da representação documental, não era valorizado. Através desse trabalho intelectual, os melhores termos, ou seja, os melhores elementos de metadados são selecionados para prover ao leitor (pesquisador) uma resposta eficiente, lendo-se eficiência no sentido de encontrar a obra (digital ou não) que possa responder à necessidade com maior precisão e atinência.

Ao abordar os metadados e sua associação a determinado item eletrônico, pode-se fazer referência a dois tipos de possibilidades na sua implementação: metadados *internos* e os metadados *externos*. Os metadados internos estão literalmente dentro do arquivo digital, junto à codificação do mesmo. Um exemplo clássico desse tipo de abordagem são os arquivos de informação em áudio (o formato digital MP3): o arquivo que carrega a codificação para a tradução do áudio traz consigo alguns elementos de metadados no padrão ID3[17]. Já os metadados externos são arquivos textuais que podem ou não utilizar algum padrão notacional a fim de descrever um arquivo ou recurso (elementos de *hardware*, por exemplo) no contexto digital. Essa técnica tem sido, atualmente, o caminho no desenvolvimento da *web* semântica de Tim Berners Lee[18]: metadados utilizando os elementos DC que utilizam a notação (escrita) no formato RDF[19] pretendem descrever

[15] Estruturas de organização informacional baseado em linhas e colunas.
[16] Capacidade de ligar as informações a fim de extrair estruturas complexas no ambiente do banco de dados.
[17] Padrão de metadados utilizado pelo formato digital em arquivos de áudio MP3.
[18] Físico Britânico, cientista da computação e professor do Massachussetts Institute of Technology. É creditado a ele o titulo de inventor da *World Wide Web*.
[19] *Resource Definition Language* – Linguagem criada para representar a informação com foco na *Web* Semântica.

recursos (arquivos e *hardware*) na Internet. Esse mesmo raciocínio pode ser implementado em uma realidade específica e de menor complexidade, como um acervo local, específico, mas nota-se claramente o uso e a importância dos metadados no processo de descrição e recuperação da informação digital.

E, embora sejam feitas considerações sobre padrões de metadados definidos, a dinâmica de criação de um padrão específico para determinada realidade/organização é plenamente possível, não havendo obrigatoriedade de se adotar um modelo fechado. Foulonneau e Riley descrevem essa possibilidade:

> Não há um único padrão de metadados que pode ser considerado o melhor para todo tipo de projeto digital. Projetar um padrão adequado começa com a análise da fonte de dados para cada elemento de metadados, em seguida, avaliar a organização desses elementos em grupos lógicos com base em como e quando serão criados. (FOULONNEAU; RILEY, 2008, p. 43).

Pode-se perceber, portanto um crescente número de padrões que, possuem uma base comum de elementos metadados originados de um padrão, como o Dublin Core, porém acrescido de suas especificidades, como exemplo MTD-BR, ETD-MS, Adobe XMP entre outros.

2.6.2 Machine-Readable Cataloging - MARC 21

Ao se refletir sobre metadados, um padrão que merece destaque é o *Machine-Readable Cataloging Record* – MARC, que passou por várias versões e atualmente é um modelo grandemente usado por bibliotecas do mundo todo. Nele, é utilizado um número relativamente elevado de elementos e suas especificações, a fim de abranger um grande número de possibilidades descritivas. Essa característica, que possui um caráter positivo em relação ao poder de representação, apresentava uma dificuldade em sua manipulação, visto que, para manusear seus elementos e especificidades, era preciso um domínio da linguagem notacional, ou seja: o MARC não é um mecanismo intuitivo, e essa característica foi levada em consideração nos avanços de metadados para a geração digital.

Ainda que o MARC pudesse ter encontrado um grande facilitador[20] no processo computacional, havia uma sobrecarga de informação descritiva, em se tratando de documentos digitais. O MARC era muito poderoso para uma nova realidade, e seu poder traria um entrave na dinâmica informacional. Diante disso, é criado em Dublin, Ohio, um

[20] Pois sua complexidade nesse ambiente é minimizada com o auxílio de modernas práticas de interação através de elementos gráficos que pretendem conduzir o usuário no preenchimento de seus elementos

padrão de metadados denominado *Dublin Core*[21] (como veremos mais adiante neste trabalho). Esse padrão pretendia utilizar um "*core*" (essência) de elementos a fim de formar metadados que pudessem ser, ao mesmo tempo, poucos e com grande potencial descritor. Houve um grande avanço com essa iniciativa, e o sucesso desse trabalho resulta na consideração de que o *Dublin Core* é hoje referência mundial em padrão de metadados digitais, sendo a instituição *Dublin Core® Metadata Initiative* (DCMI) a principal validadora desse padrão. A criação de um novo padrão de metadados, resumido, era necessária, pois, no ambiente digital, novos elementos passaram a coexistir com as obras literárias e com outras possibilidades para sua descrição, sendo que texto, áudio e vídeo passaram a formar um novo conjunto: a multimídia.

Cíntia Azevedo Lourenço exemplifica essa dualidade de padrões através do conflito entre a adoção de uma estrutura mais abrangente ou essencial:

> Os minimalistas defendem o uso de um padrão mais simples, com apenas os elementos principais para uma boa recuperação. Já os estruturalistas defendem o uso do formato de descrição completo do padrão MARC para identificação de documentos na *web*. Contudo, tanto os minimalistas quanto os estruturalistas tentam adotar um padrão clássico da biblioteconomia para a estruturação de um padrão de metadado eficiente para a organização de bibliotecas e arquivos digitais. (LOURENÇO, 2007, p. 77).

Essa reflexão é importante, porque reflete a necessidade da adoção de um padrão comum – independente do número de elementos – que possa ser usado para identificar um item digital.

O formato MARC foi criado para permitir o registro dos dados catalográficos em formato eletrônico, de modo que o computador possa armazenar, processar e recuperar as informações. Esse formato é fruto de pesquisas da *Library of Congress* (LC) em conjunto com bibliotecas universitárias americanas, com o objetivo de especificar mecanismos que pudessem ajudar na automação de suas atividades. As primeiras especificações ocorreram na década de 60, sendo seu início nomeado "MARC *Pilot Project*". Ainda nessa década, o formato tornou-se operacional e começou a ser usado pela LC.

Com o passar do tempo, o formato passou por diversas transformações e atualizações, o que é considerado natural para um modelo que pretende abranger uma correspondência mundial. Um fato a ser considerado diz respeito ao formato ter sido regionalizado, criando-se formatos variados, que atendam a necessidades nacionais. Esse é um ponto que merece destaque, pois a intenção de padronizar determinado volume informacional consiste em estabelecer regras universais que possam ser seguidas. Os formatos regionalizados também permitem a integração das informações com a

[21] Padrão de metadados criado pela DCMI (*Dublin Core Metadata Initiative*) a fim de prover um vocabulário aberto e interoperável na descrição de recursos eletrônicos.

consequente interoperabilidade entre sistemas, com exceção dos formatos INTERMARC[22] e UNIMARC[23], usados pela biblioteca da França, pois sua designação de conteúdo próprio não é compatível com o MARC 21.

O MARC 21 nasceu da fusão entre os formatos USMARC[24] (LC) e CANADIAN MARC[25] (Biblioteca Nacional do Canadá). Hoje, no Brasil, é usado o MARC 21.

Figura 7 - Exemplo de informações em formato MARC

```
000      01691cpcaa2200289 a 450
001      468 3524
005      20070511132339.0
008      961118i1918 1970ctu eng d
035 __   $9 FNN178 4YL
040 __   $a CtY-D $c CtY-D $e dacs
090 __   $a RG $b 61
245 10   $a Issues of Peace and War Pamphlet Collection, $f 1818-1970 (inclusive)
300 __   $a 9.5 $f linear feet (18 boxes)
351 __   $a Organized into two series: I. Organizations. II. Topical Material.
520 __   $a Material documents the issues related to the Second World War, with documents from as early as 1818
           and as late as 1970 are included. The collection includes material from more than 170 organizations as
           well as topically arranged materials. The collection is particularly valuable for its focus on the contribu-
           tions of American religious organizations and church leaders to the discussion and action surrounding
           such issues as conscientious objection, civilian public service camps, military training , disarmament, and
           reconstruction.
546 __   $a Chiefly in English.
506 __   $a This collection is open for research.
555 0_   $a Finding aid available in repository and on internet.
524 __   $a Issues of Peace and War Pamphlet Collection (RG 61). Special Collections, Yale Divinity School Library.
610 20   $a Fellowship of Reconciliation (U.S.)
650 _0   $a Peace $x Societies, etc.
650 _0   $a Church and social problems.
650 _0   $a World War, 1939-1945.
650 _0   $a War
650 _0   $a Conscientious objection.
852 __   $a Special Collections, $b Yale Divinity School Library, $e 409 Prospect St., New Haven, CT 06511
856 42   $3 Finding aid $u http://webtext.library.yale.edu/xml2html/divinity.061.con.html
```

Fonte: <http://www.ub.edu/bid/21/estiv2.htm>

Independentemente de sua versão, o formato foi criado com o objetivo de estruturar a informação de modo a permitir que ela seja legível por *software*. Isso significa que uma máquina computacional é capaz ler e interpretar a informação armazenada no registro catalográfico através de um *software* adequado.

Na catalogação apoiada por computador, o uso do formato MARC é imprescindível para a entrada dos elementos de dados no subsistema de catalogação do *software* adotado pela agência catalogadora. A fim de estabelecer cooperação de registros e interoperabilidade entre sistemas, esse formato obteve grande aceitação mundial, havendo, naturalmente, algumas exceções.

[22] Formato utilizado pela Biblioteca Nacional da França.
[23] Formato criado pela IFLA (*International Federation of Libraries Association*) em 1977.
[24] Formato americano.
[25] Formato canadense.

Essa reflexão sobre o MARC é importante, porque outros formatos foram criados, posteriormente, sob sua forte influência. Um exemplo disso é o formato *Dublin Core*, enfocado a seguir.

2.6.3 Dublin Core

O *Dublin Core Metadata Element Set* (DCMES) (2011), ou simplesmente *Dublin Core* – DC, é um conjunto mínimo de metadados, criado para descrever documentos eletrônicos. Criado originalmente para a descrição de documentos eletrônicos, seu desenvolvimento foi originado na *Online Computer Library Center* (OCLC) (2000).

Esse padrão descreve um conjunto de campos essenciais, a fim de permitir a identificação de um determinado arquivo/recurso com interoperabilidade[26] e transparência. Seus elementos expressam metadados encontrados em diversos outros padrões mais complexos, permitindo uma interoperabilidade quando usado, pois este conjunto torna possível expressar informações comuns encontradas em outros formatos.

O DC pretendia estabelecer os mais importantes descritores, inicialmente, através de treze elementos. A simplicidade era alcançada sem perder o poder descritivo e a padronização precisava ser considerada para um passo ainda maior, a automação do processo de pesquisa computacional. Hoje, no período de construção e publicação deste trabalho, esse padrão conta com quinze elementos.

Vários dispositivos de busca e indexação na Internet, conhecidos como robôs, fazem uso do DC para indexar informações e, posteriormente, disponibilizá-las para serviços de busca. Deve-se reconhecer que esse padrão foi criado com um forte direcionamento para a solução de pesquisa de itens (arquivos/recursos) publicados na Internet. Implementar uma indexação automática na Internet é uma tarefa complexa, considerando que, além da indexação, os resultados são, em sua extensa maioria, um conjunto imenso de referências. Se o arquivo/recurso contém referências a respeito de um campo de estudo específico, o vocabulário pode dificultar ainda mais todo o processo de indexação e recuperação.

A intenção do padrão DC[27], diferente de outros padrões, é tornar-se genérico e, ao mesmo tempo, ser dotado de grande poder de descrição, o que o torna acessível a uma grande variedade de aplicações.

A *World Wide Web Consortium* (W3C) (2012), instituição internacional liderada por Tim Berners Lee, dedicada a estudar e publicar padrões para Internet, recomenda utilizar esse padrão de descritores em conjunto com a linguagem RDF em *softwares* que, através dessas tecnologias, permitirão aos usuários finais descrever as informações que

[26] Característica que permite o acesso à informação independente de fabricantes de *software* e *hardware*.
[27] Os elementos do DC estão descritos no Apêndice A deste trabalho.

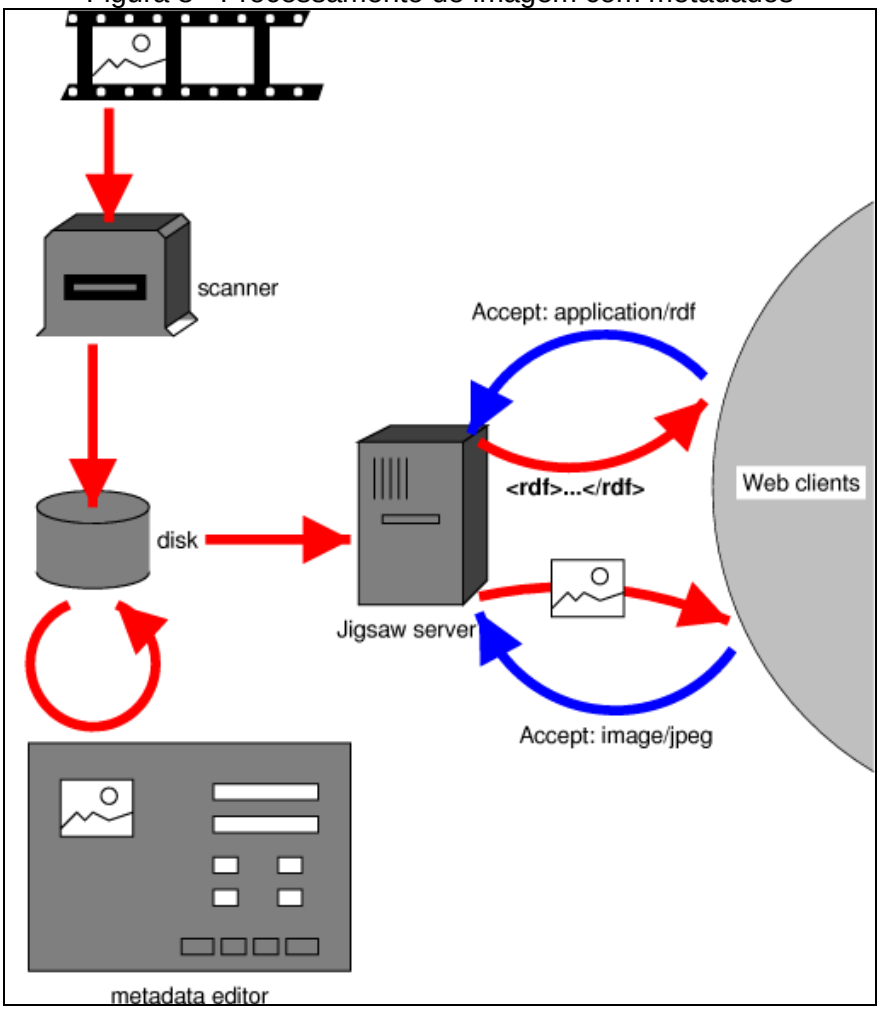
poderão ser usadas por sistemas de recuperação da informação (SRI) em diversos ambientes, sejam eles locais ou através da Internet.

Ao abordar essa teoria, pode-se visualizar um exemplo prático publicado na W3C (LAFON; BOS, 2002) utilizando para isso, o tipo de documento “imagem”. O contexto expressa o seguinte raciocínio:

Os autores do projeto possuem muitas fotos e gostariam de recuperá-las de maneira mais objetiva.

Usando as recomendações da W3C, foi criado o seguinte processo (figura 8):

Figura 8 - Processamento de imagem com metadados



Fonte: (LAFON; BOS, 2002)

a) as figuras são digitalizadas ou são produzidas já no formato eletrônico através de câmeras digitais;
b) após o processo de digitalização (para figuras em formato físico), os arquivos são gravados em disco rígido e serão documentados;

c) para realizar o processo de inserção dos metadados, utiliza-se um *software* (Figura 9) que disponibiliza o esquema DC de metadados, permitindo a gravação das informações embutidas no arquivo jpg[28] ou permitindo gerar uma estrutura externa ao arquivo em formato RDF (como veremos adiante) usando o esquema DC em sintaxe XML;

d) esses arquivos são gravados em um servidor de arquivos que receberá requisições de pesquisa e, ao pesquisar nos metadados, responderá com a(s) imagem(ns) que atendam à solicitação.

Figura 9 - *Software* de atribuição de metadados em imagens

Fonte: Elaborado pelo autor

Desse modo, percebe-se como o padrão DC pode ser utilizado na prática, através de *softwares* e procedimentos que documentam o recurso eletrônico, possibilitando sua recuperação posteriormente. Todo esse movimento depende, contudo, dos *softwares* disponíveis para o usuário final. Os programas populares precisam oferecer ferramentas práticas ou interfaces rápidas e amigáveis, além da adoção coerente de padrões (esquemas) que possibilitem os serviços de recuperação.

Um resumo dos elementos do padrão *Dublin Core* é descrito no Quadro 2, e uma descrição detalhada encontra-se disponível no apêndice A deste trabalho.

[28] É um método usado para comprimir imagens em diferentes níveis de redução de tamanho de armazenamento. Consequentemente, é um método que afeta a qualidade da imagem.

Quadro 2 - Resumo dos elementos do padrão *Dublin Core*

title (título)	O nome dado a um determinado recurso.
creator (criador)	A entidade primária responsável (em primeira instância) pela criação do conteúdo do recurso.
subject (assunto)	O tópico do conteúdo do recurso.
description (descrição)	Uma descrição do conteúdo do recurso.
publisher (editor)	A entidade responsável por tornar o recurso acessível.
contributor (colaborador)	A entidade responsável por registrar os colaboradores do conteúdo do recurso.
date (data)	A data associada ao evento de ciclo de vida do recurso.
type (tipo)	A natureza ou gênero do conteúdo do recurso.
format (formato)	A manifestação física ou digital do recurso.
identifier (identificador de recurso)	Uma referência não ambígua (única) a fim de especificar o recurso em um determinado contexto.
source (fonte)	A referência da qual o recurso foi derivado.
language (língua)	A linguagem intelectual do conteúdo do recurso.
relation (relação)	Permite documentar a referência a um recurso relacionado.
coverage (cobertura)	A descrição da extensão do escopo do conteúdo do recurso.
rights (direitos)	A informação dos direitos sobre o recurso.
audience (audiência)	Permite qualificar a pretensão do público de usuários para o recurso.

Fonte: Elaborador pelo autor

Rosetto categoriza esses elementos em 3 grandes grupos:

Quadro 3 - Categorias dos elementos de metadados do formato *Dublin Core*

Conteúdo	**Propriedade Intelectual**	**Características físicas**
Título	Criador	Data
Assunto	Editor	Tipo (de recurso)
Descrição	Contribuinte	Formato (suporte físico)
Fonte	Direitos	Identificação (local)
Idioma		
Relação		
Cobertura (extensão)		

Fonte: (ROSETTO, 2003)

Uma iniciativa que aponta para o movimento de adoção deste padrão no mercado é o exemplo do fabricante de *softwares* Adobe. A empresa adotou em seus *softwares* o padrão XMP[29] (ADOBE SYSTEMS, 2012) de metadados baseado em DC. Esta iniciativa constitui um excelente exemplo de como a indústria pode fomentar a documentação de recursos de maneira mais efetiva. Adiante, pode-se ver o padrão MTD-BR para documentação de documentos científicos brasileiros (teses e dissertações), modelo também baseado em *Dublin Core*.

2.6.4 Biblioteca Digital de Teses e Dissertações – padrão MTD-BR

No Brasil, a Biblioteca Digital de Teses e Dissertações – BDTD (INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA E TECNOLOGIA, 2012) iniciou um projeto de integração das publicações eletrônicas e registro bibliográfico das teses e dissertações das Instituições de Ensino Superior (IES). O projeto esta sendo realizado através de uma metodologia que envolve a adoção de um padrão de metadados e uma plataforma tecnológica para a transmissão das informações que estão nas universidades para o servidor da BDTD. Esse projeto tem sido de grande relevância no círculo acadêmico.

O modelo adotado foi inspirado na *Networked Digital Library of Thesis and Dissertation*[30] (NDLTD), iniciativa que construiu uma biblioteca digital de teses e dissertações em âmbito internacional, liderado pela *Virginia Tech*, Universidade de Michigan.

No modelo adotado no Brasil pelo IBICT – Instituto Brasileiro de Informação em Ciência e Tecnologia, a biblioteca é distribuída, sendo as Instituições de Ensino Superior geradoras dos metadados e do conteúdo, e o IBICT é responsável pela 'colheita' (*harvest*[31]) e exposição da informação através de uma interface única, simplificada e de fácil acesso.

Nesse modelo, o IBICT é considerado o provedor de serviços nacional e o provedor de dados a nível internacional. Um modelo simplificado desse processo pode ser visto na figura 10.

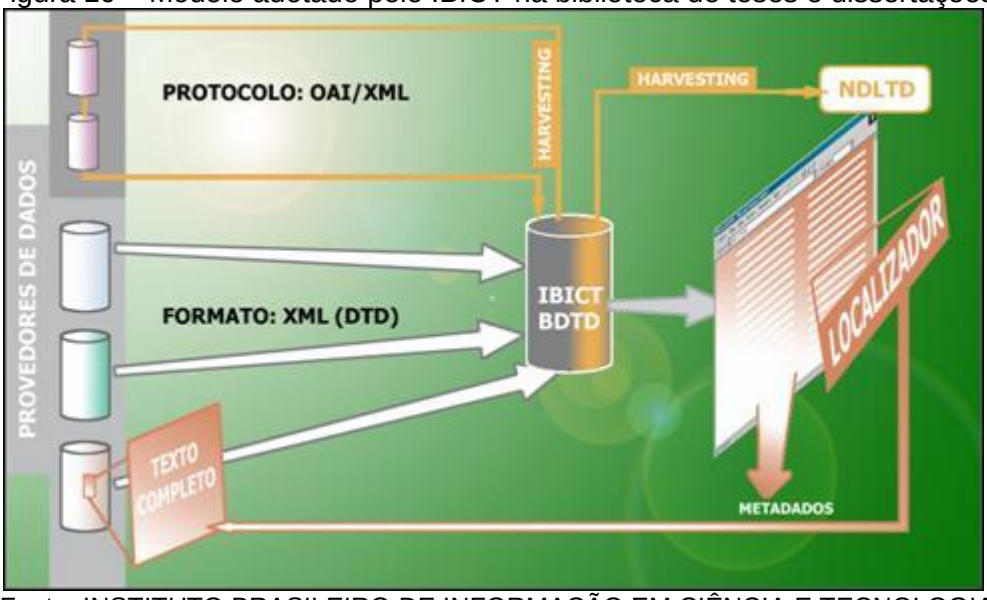
---

[29] *Extensible Metadata Plataform* é o padrão de metadados adotado pela Adobe em diversos *softwares* de sua plataforma.

[30] Biblioteca Digital de Teses e Dissertações em Rede (VIRGINIA TECH, 2012).

[31] *Harvest* é um processo automatizado no qual provedores de serviços ou agregadores coletam metadados

Figura 10 – Modelo adotado pelo IBICT na biblioteca de teses e dissertações



Fonte: INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA E TECNOLOGIA, 2006

A grande vantagem deste modelo é permitir que qualquer usuário que tenha acesso à Internet acesse um único endereço (portal) para realizar pesquisas nessa base de dados, que reúne as referências ao documento original, na sua instituição publicadora, através da leitura de metadados produzidos anteriormente. Ou seja, o usuário não precisará procurar em cada instituição pelo documento que lhe atenda o interesse, pois suas referências estão reunidas em um único banco de dados.

Adotando-se um princípio de independência entre a gestão local das Instituições de Ensino Superior e a BDTD, porém, observa-se que, para a publicação do acervo da instituição de ensino na BDTD, é necessária a utilização de padrões de metadados e de transferência desses elementos, para que todo o processo possa funcionar adequadamente.

Southwick (2003) enfatiza que, para a definição desse padrão os seguintes fatores foram levados em consideração:

a) estudar experiências existentes no Brasil e no exterior de desenvolvimento de bibliotecas digitais de teses e dissertações;
b) desenvolver, em cooperação com membros da comunidade, o modelo para o sistema da BDTD;
c) definir padrões de metadados e tecnologias a serem utilizados pelo sistema da BDTD;
d) absorver e adaptar as tecnologias a serem utilizadas na implementação do modelo;

e) desenvolver um sistema de publicação eletrônica de teses e dissertações para atender àquelas IES que não possuem sistema automatizado para implantar suas bibliotecas digitais;
f) desenvolver procedimento automatizado para permitir a integração dos catálogos de dezessete universidades brasileiras com as bibliotecas digitais locais e com a BDTD nacional;
g) difundir os padrões e tecnologias adotados na BDTD e dar assistência técnica aos potenciais parceiros na sua implantação.

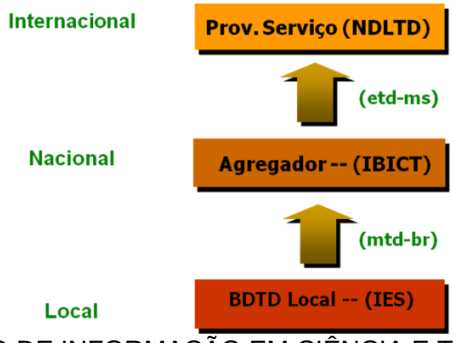
Ainda segundo Southwick (2003), esse modelo apresenta os seguintes conceitos:

a) *Metadados* – a definição de metadados é "dado sobre dado". No âmbito do sistema BDTD, utilizam-se três padrões de metadados relacionados a teses e dissertações. O foco principal desses metadados é a descrição do objeto digital (tese ou dissertação) e sua localização na Internet.
b) *Provedor de dados* – entidade que administra sistemas que mantêm repositório de dados e suportam o Protocolo OAI-PMH[32] como meio de expor metadados para serem coletados por provedores de serviço ou agregadores.
c) *Provedor de serviço* – entidade que usa metadados coletados automaticamente dos provedores de dados via protocolo OAI-PMH, como base para oferecer produtos e serviços de valor agregado.
d) *Agregador* – entidade que coleta metadados, construindo repositórios centralizados com eles, e atua como provedor de dados para outros provedores de serviço. Agregador, portanto, exerce tanto o papel de provedor de dados como de provedor de serviço.
e) *Coleta automática de metadados (metadata harvesting)* – é um processo automatizado para coleta de metadados dos repositórios dos provedores de dados, por meio do uso do protocolo OAI-PMH.
f) *Protocolo OAI-PMH*: *Open Archives Iniciative Protocol for Metadata Harvesting* (2002) – Esse protocolo opera sobre o protocolo *http*. Os provedores de serviço enviam solicitações de metadados aos provedores de dados. Estes respondem com metadados estruturados em registros XML, obedecendo a um padrão de metadados. O protocolo OAI-PMH provê um modelo de interoperabilidade baseado no processo de coleta automática de metadados (*metadata harvesting*).

[32] Pode ser traduzido como "protocolo aberto de arquivos para coleta de metadados".

Uma visão esquemática desse modelo pode ser vista através da Figura 11, que mostra como as Instituições de Ensino Superior organizam seus acervos em níveis locais e disponibilizam os metadados no padrão MTD-BR para o IBICT.

Figura 11 – Esquema geral de funcionamento do modelo IBICT



Fonte: INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA E TECNOLOGIA, 2006

Este processo ocorre utilizando os seguintes padrões:

a) DC: Dublin Core (2011);
b) Mtd-br: padrão **Br**asileiro de **M**etadados para **T**eses e **D**issertações (INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA E TECNOLOGIA, 2012);
c) Etd-ms: ***E****lectronic* ***T****hesis and* ***D****issertation* ***M****etadata* ***S****tandard*, padrão adotado pela NDLTD (Biblioteca de Teses e Dissertações internacional) (NETWORKED DIGITAL LIBRARY OF THESES AND DISSERTATIONS, 2012).

Southwick relaciona esses três padrões da seguinte forma:

> O padrão de metadados *Dublin Core* é exigido pela Iniciativa de Arquivos Abertos como o conjunto mínimo de metadados a ser exposto pelos repositórios aderentes a essa iniciativa. Os elementos do padrão de metadados *Dublin Core* são comuns aos dois outros padrões usados pela BDTD. O padrão etd-ms, por sua vez, é um subconjunto do padrão mtd-br. (SOUTHWICK, 2003).

Ainda segundo Southwick (2003), “o padrão *etd-ms* inclui todos os elementos do padrão DC e, adicionalmente, inclui elementos específicos para teses e dissertações.” Esses elementos adicionais são:

a) Titulação – nome do grau associado com a tese ou dissertação, como aparece no documento. Exemplo: Mestre em Pesquisa Operacional;
b) Grau – nível de educação associado com o documento. Exemplo: mestre, doutor;

c) Disciplina – área de estudo do conteúdo intelectual do documento. Usualmente, indica-se o nome do programa de pós-graduação ou departamento;
d) Instituição que abriga o programa de pós-graduação.

Como o padrão MTD-BR é uma extensão do padrão ETD-MS, ele possui todos os metadados do padrão DC, incluindo os quatro elementos do ETD-MS e, segundo Southwick (2003), é acrescido dos seguintes itens, denominados *classes de metadados*:

a) Metadados de gestão do registro – nessa classe, são incluídos metadados para a identificação única do registro, identificação da instituição cooperante provedora do registro de metadados, especificação do tipo de conteúdo do registro e metadados relacionados a restrições de uso do objeto digital descrito;
b) Metadados de descrição da tese ou dissertação – esses metadados servem para descrever a tese ou dissertação. São os metadados de descrição bibliográfica para teses e dissertações, tais como título, autor, resumo;
c) Metadados para a identificação de pessoas – para pessoas, tais como autor e contribuidores (membros da banca), são especificados metadados que as identifiquem, sempre que possível, por meio do metadado CPF exclusivamente no Brasil. Para estrangeiros, este metadado é opcional;
d) Metadados para a identificação de instituições. – várias instituições podem estar direta ou indiretamente relacionadas com uma tese ou dissertação. As instituições identificadas no mtd-br são: a instituição que abriga o programa de pós-graduação; afiliações de autor e contribuidores (membros da banca); agência de fomento que financiou integral ou parcialmente o trabalho de pesquisa que deu origem à tese ou dissertação;
e) Metadados de ligação – esses metadados servem para referenciar, por meio de endereços eletrônicos, objetos digitais ou páginas *web* relacionadas à tese ou dissertação descrita.

Resumindo-se as alterações, tem-se o seguinte:

a) *Dublin Core* foi a base dos metadados;
b) o ETD-MS acresceu 4 elementos adicionais ao DC;
c) o MTD-BR acresceu 5 classes de elementos adicionais ao etd-ms.

Essa abordagem teve como fator motivacional a compatibilidade com outros bancos de dados, como a Plataforma Lattes[33]. Um detalhamento do padrão MTD-BR encontra-se na página do IBICT na Internet.

Notou-se que, durante todo o processo referido acima, padrões de escrita e gravação permanente dos metadados são considerados. Descrevemos a seguir algumas tecnologias fundamentais para a implementação efetiva desse e de outros modelos, usando XML e RDF, visando a uma arquitetura aberta e à interoperabilidade entre tecnologias.

2.7 Esquemas para gravação de metadados

Metadados, enquanto estruturas informacionais, precisam ser associadas a um determinado item digital e gravadas de modo permanente, a fim de permitir a recuperação em momento oportuno. Essa informação pode estar inserida no mesmo arquivo ou gravada em uma estrutura independente (outro arquivo), que deve usar uma estrutura sintática a fim de estabelecer um padrão comum na leitura humana e computacional, como é visto adiante.

2.7.1 *eXtensible Markup Language* – XML

Um problema comum na apresentação de uma informação que pretende ser universal é sua forma de representação. A *eXtensible Markup Language* foi uma resposta simples e inteligente que impactou na maneira como a tecnologia aborda problemas relativos à interoperabilidade e publicação (universalização) de conteúdo.

A XML tem seu início na década de 1960. Um grupo de cientistas da *International Business Machines* (IBM) criou uma linguagem de computação baseada em marcação voltada para o processamento de documentos. Surge a especificação da *Standart Generalized Markup Language* (SGML) que, infelizmente, por causa da sua complexidade, não foi adotada pelo mercado. Havia necessidade de uma tecnologia igualmente rica em recursos, de fácil entendimento, simples e dinâmica, pois a *World Wide Web* ainda estava no início e precisava de uma linguagem de marcação leve, aberta, flexível e simples. Foi criado, então, o *HyperText Markup Language* (HTML), como um resumo da SGML, o que gerou muitos avanços no processo de criação e divulgação da Internet, porém essa linguagem apresentava algumas limitações que precisavam ser superadas. O número limitado de marcadores (*tags*) e sua estrutura técnica apontavam para a necessidade de mudanças. Um problema da HTML é que ela foi escrita para a camada de apresentação, não sendo sua codificação transparente para o usuário.

[33] A Plataforma Lattes representa a experiência do CNPq na integração de bases de dados de Currículos, de Grupos de pesquisa e de Instituições em um único Sistema de Informações. (CONSELHO NACIONAL DE DESENVOLVIMENTO CIENTÍFICO E TECNOLÓGICO, 2012).

A linguagem obedece a um padrão muito simples, herança das linguagens de marcações anteriores, nas quais o símbolo *tag* é composto pelos caracteres maior e menor:

<escola> UFMG </escola>

Deitel (2003) afirma que é necessário um *software* chamado *analisador sintático* (*parser*) de XML (ou um processador XML) para processar um documento XML, o qual lê o documento, verifica sua sintaxe, relata quaisquer erros e permite, via programas, acesso ao conteúdo do documento.

Neste projeto, pretende-se utilizar o analisador sintático nativo do sistema *Windows* da *Microsoft*, o *msxml*.

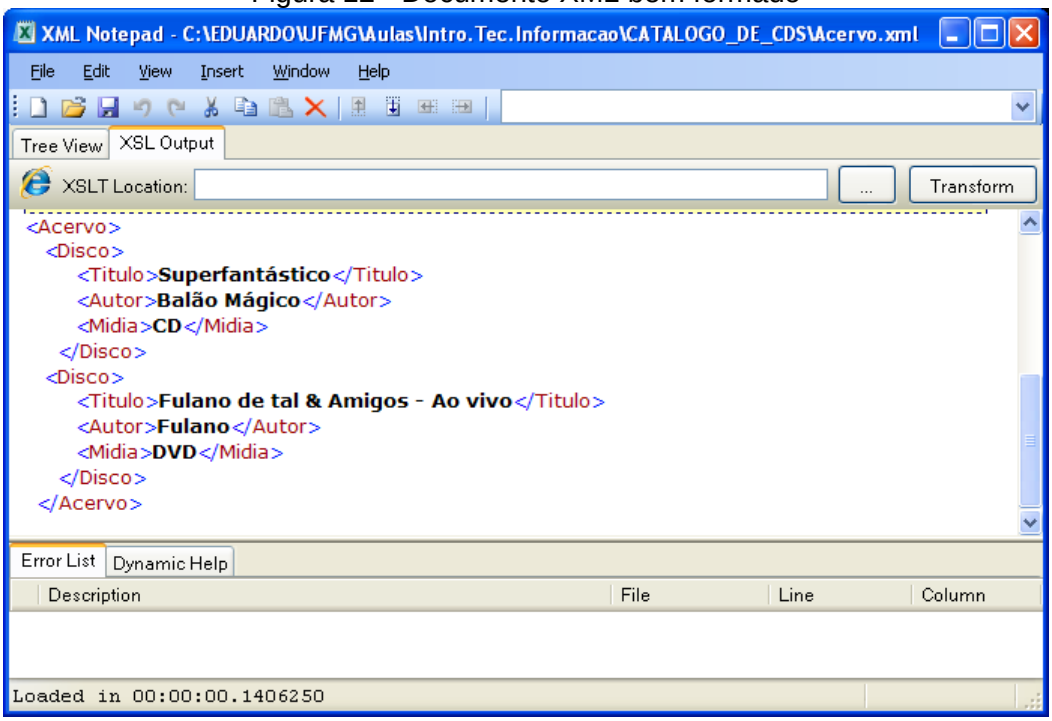
É importante notar que o documento XML deve obrigatoriamente ser *bem formado* para ser utilizado. Um documento válido é denominado um documento *bem formado*, e possui as seguintes características, segundo Deitel (2003):

a) possuir um único elemento como nó raiz;
b) apresentar uma marca de abertura e de finalização para cada elemento;
c) apresentar as marcas corretamente aninhadas;
d) mostrar os valores de atributos entre aspas.

A XML também diferencia entre caixa alta (maiúsculas) e caixa baixa (minúsculas), de modo que as nomenclaturas de elementos usados no arquivo devem ser padronizadas.

Na Figura 12, pode-se observar um arquivo XML bem formado através do *software* de edição XML, *Notepad*. (MICROSOFT, 2007).

Figura 12 - Documento XML bem formado



Fonte: Autor

A XML atua como linguagem (forma de comunicação) livre, aberta, sem restrições em sua codificação, pois é gravada em texto puro. É extremamente versátil, permitindo o desenvolvedor ser capaz de estabelecer suas marcações para que os demais *softwares* possam recuperá-la. E, nesse contexto, outras linguagens fazem uso da XML, a fim de usar suas características e sintaxe, como é o exemplo da RDF.

2.7.2 *Resource Definition Framework* – RDF

Como foi visto anteriormente, as linguagens de marcação têm um importante papel no processo de organização informacional no espaço digital. Diante da crescente necessidade de tornar a recuperação das informações de seu acervo (que apresenta grande número de arquivos) mais eficiente, foi criado o modelo padrão *Resource Definition Framework* (RDF). A RDF é recomendada pela W3C para representar a informação na *Web* (WORLD WIDE WEB CONSORTIUM, 2012b), e, segundo Deitel (2003, p. 652), a RDF é uma linguagem baseada em XML que descreve informações contidas em um recurso, seja uma página *web*, um *site* completo ou qualquer item eletrônico que contenha informação.

Essa capacidade de realizar a descrição do recurso, ou seja, implementar um modelo de metadados, possui uma enorme amplitude de utilização, podendo ser utilizada

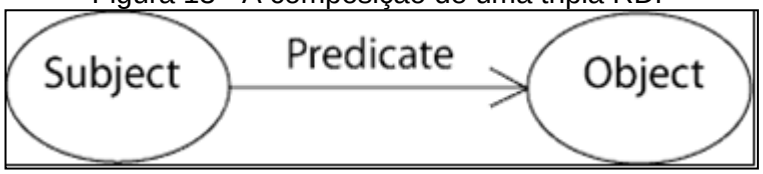
por buscadores robôs[34] na Internet e por *softwares* proprietários, para extrair de forma padronizada informações sobre o recurso. A W3C afirma que a RDF é a tecnologia base para a realização da ligação entre os dados, fazendo com que o conteúdo espalhado e sem conexão seja interligado, constituindo, então, a base para a implementação efetiva da *Web* Semântica.

É importante entender que a RDF foi criada para que as máquinas possam processar a informação que será disponibilizada para os usuários. A W3C descreve essa ligação como o coração da *Web* Semântica: a possibilidade de integração em larga escala, raciocínio e informação (descritivas) sobre a própria *Web*, ou seja, a possibilidade de identificar recursos através de identificadores e valores associados.

O princípio é simples e baseia-se no conceito composto de três elementos, denominado tripla:

a) Sujeito (objeto a ser descrito);
b) Predicado (declaração);
c) Objeto (identificador).

Figura 13 - A composição de uma tripla RDF



Fonte: (BECKETT, 2004)

A título de exemplo, a declaração:

http://www.xyz.com.br	foi criado por	Fulano xyz da Silva

representa a tripla RDF da seguinte forma:

Sujeito: http://www.xyz.com.br
Predicado: foi criado por
Objeto: Fulano xyz da Silva

[34] Programas que vasculham páginas e arquivos (recursos) na Internet a fim de coletar informações servindo a grandes bases de dados que, posteriormente, estabelecem parâmetros de indexação do conteúdo coletado.

A linguagem natural expressa a comunicação humana. A RDF foi criada para permitir o processamento de máquinas e humanos através de suas declarações. Usando a linguagem para expressar o exemplo anterior, haveria a seguinte sintaxe:

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 xmlns:dc="http://purl.org/dc/elements/1.1/">
<rdf:Description rdf:about="http://www.xyz.com.br">
<dc:creator>Fulano xyz da Silva</dc:creator>
</rdf:Description>
</rdf:RDF>
```

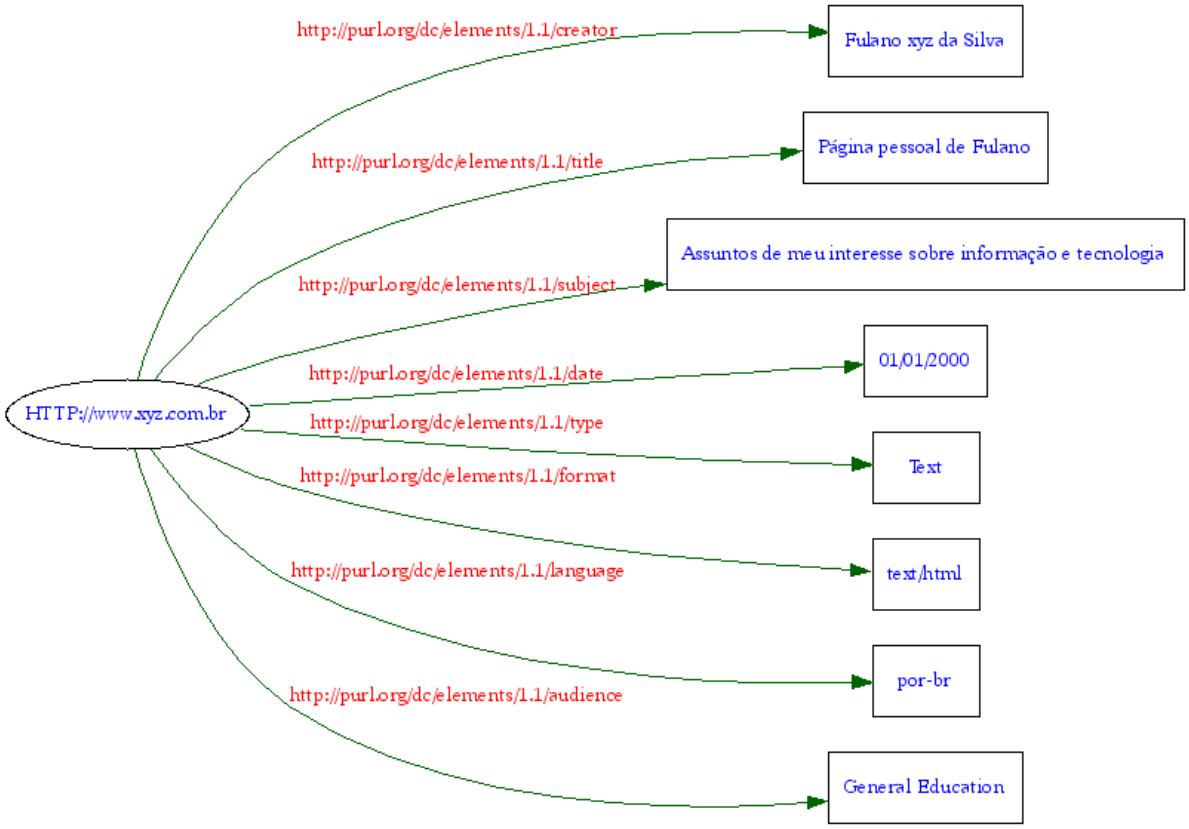
Ao estender o exemplo acima, acrescentando mais declarações ao sujeito (http://www.xyz.com.br), o arquivo RDF se torna ainda mais descritivo:

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 xmlns:dc="http://purl.org/dc/elements/1.1/">
<rdf:Description rdf:about="HTTP://www.xyz.com.br">
<dc:creator>Fulano xyz da Silva</dc:creator>
<dc:title>Página pessoal de Fulano</dc:title>
<dc:subject>Assuntos de meu interesse sobre informação e tecnologia</dc:subject>
<dc:date>01/01/2000</dc:date>
<dc:type>Text</dc:type>
<dc:format>text/html</dc:format>
<dc:language>por-br</dc:language>
<dc:audience>General Education</dc:audience>
</rdf:Description>
</rdf:RDF>
```

A fim de fomentar essa tecnologia, a W3C disponibiliza um *software* validador[35] via *Web* (figura 14) para avaliar o código RDF, através do qual pode-se ver se a sintaxe (escrita) está correta ou não. O serviço também exibe uma visão gráfica que ajuda a entender melhor como as relações de triplas funcionam dentro da linguagem.

[35] *Software* destinado a processar o código e exibir um relatório com erros ou apresentação final do propósito do código computacional.

Figura 14 - Validador de código RDF e sua saída gráfica

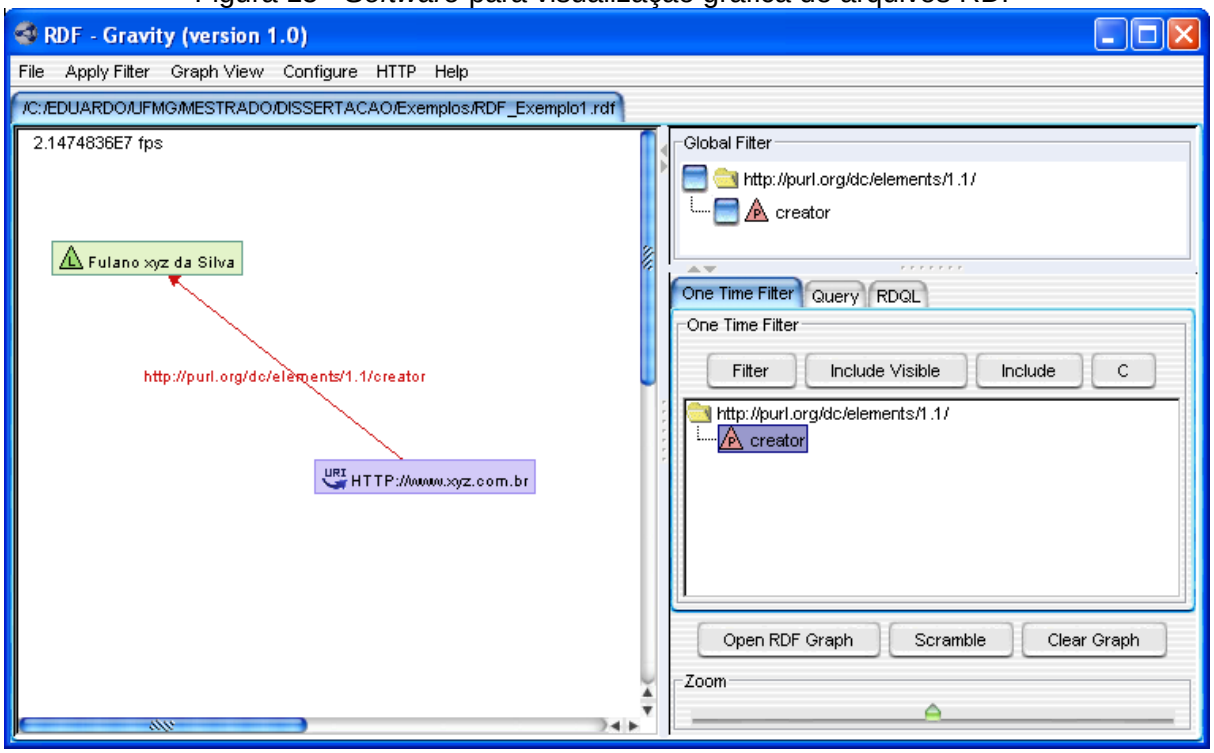


Fonte: <http://www.w3.org/RDF/Validator/ >

Observa-se, nesse exemplo, que os elementos usados para estabelecer os metadados foram extraídos do padrão *Dublin Core* (mas poderia ser qualquer outro padrão), de modo que diversas tecnologias trabalham juntas a fim de formar um resultado.

Outro aplicativo que permite observar a linguagem RDF de forma gráfica é o *software* RDF *Graph VIsualization Tool* (figura 15).

Figura 15 - *Software* para visualização gráfica de arquivos RDF



Fonte: <http://semweb.salzburgresearch.at/apps/rdf-gravity/>

Observa-se que o padrão RDF é escrito através da sintaxe em XML e pode ser gravado em arquivos independentes ou de forma embutida no próprio arquivo, como viu-se no exemplo que utilizava imagens (fotos) da W3C. Ao serem lidos e processados, esses metadados expõem a informação estruturada ao leitor, humano ou computacional.

Com a finalidade de tornar esses documentos passíveis de processamento para máquinas, e não apenas para humanos, a função da *web* semântica está em relacionar itens e informações de modo a encontrá-las de forma otimizada. O modo como humanos e máquinas consomem informação é muito diferente: enquanto pessoas conseguem ler a informação em sua representação de mais alto nível, as máquinas consultam as marcações e demais codificações da página no ambiente *web* (no exemplo). Isso está mostrado na Figura 16.

Figura 16 - Comparativo das abordagens computacional e humana em uma página Web

Fonte: (ADIDA et al. , 2012)

Podem-se destacar as seguintes características neste ambiente:

a) o uso da linguagem XML para estabelecer regras de sintaxe e demais propriedades inerentes a esse padrão (bem formado);
b) os elementos para realizar a descrição são do padrão *Dublin Core*;
c) a linguagem RDF estabelece o padrão de triplas a fim de descrever o item em questão.

O potencial das linguagens de marcações alinhadas em um padrão de metadados permite que máquinas realizem inferências de maior exatidão na recuperação de informações. Essa é uma estratégia seguida pelos grandes produtores de softwares e das empresas detentoras dos motores de busca na web (*search engines*).

3 METODOLOGIA

Este capítulo aborda como a intenção da construção do protótipo foi trabalhada em seus aspectos técnicos e teóricos, fundamentados na Ciência da Informação. Projeto, expectativas, avanços e desafios estarão documentados nos tópicos que compõem o capítulo.

3.1 Tipo de pesquisa

Este estudo desenvolve práticas de natureza teórica, descritiva, exploratória e aplicada, tendo como objetivo o desenvolvimento do projeto em duas partes: teórica e prática tecnológica na construção de um *software*.

3.2 Universo de estudo

A fim de constituir uma base documental para realizar os estudos desse protótipo, foram selecionados documentos científicos para servirem de teste nos processos do *software.* Os documentos selecionados foram os seguintes:

Mestrado – Três documentos eletrônicos em formato PDF de dissertações já defendidas e cadastradas no PPGCI/UFMG da área de Organização e Uso da Informação – OUI.

Doutorado – Quatro documentos eletrônicos em formato PDF de teses já defendidas e cadastradas no PPGCI.

3.3 Projeção inicial para construção do protótipo MeTa

Para o início do funcionamento do protótipo, uma estrutura taxonômica da área de estudo do acervo foi adotada como padrão de referência no processo de indexação dos documentos. Esse vocabulário está gravado em um sistema de arquivo aberto no padrão XML (eXtensible Markup Language) e pode ser manipulado pelo usuário do *software*, que poderá adicionar, retirar termos ou criar um novo instrumento taxonômico. Esse processo é denominado importação/seleção da taxonomia e deve permitir a realização de uma indexação semi automática do documento. O processo de criação ou importação desse vocabulário é gravado nas configurações para uma utilização posterior. É importante esclarecer que esse vocabulário deve cobrir uma área específica do conhecimento, a fim de permitir uma recuperação mais assertiva posteriormente. Com essa abordagem aberta – que permite a seleção do instrumento de apoio à indexação –, pretende-se abranger um número ilimitado de áreas de conhecimento, ou seja: para a utilização dessa proposta, será

necessário que a área possua um conjunto de descritores para representar a informação. A partir do momento em que se puder associar um documento a itens desse instrumento, indexar os documentos e armazená-los com estruturas de recuperação (metadados), a possibilidade de recuperação também poderá ser realizada com maior especificidade.

Ao trabalhar o processo de indexação citado anteriormente, projetou-se um mecanismo para determinar qual descritor deve ser mais relevante para um arquivo específico. Usando-se o universo de termos da taxonomia, deverá ser utilizado um modelo de peso numérico (1 a 5) associado a partes específicas do documento. Exemplo:

Quadro 4 - Sistema de pesos para a indexação

Local do termo no texto	**Peso do termo**
Título do texto	5
Palavras chave	3
Corpo do texto	1

Fonte: Elaborado pelo autor

Este sistema de medidas visa a permitir uma configuração para que se utilizem os *n* termos mais pesados como descritores na recuperação do texto. E, embora ciente das limitações do sistema de pesquisa por termos e da repetição no documento, definiu-se que este será o método a ser usado inicialmente, a fim de viabilizar o protótipo em tempo hábil. Sabendo-se que um termo X aparece N vezes no documento, sua avaliação de importância pode ser mensurada com a seguinte fórmula:

$$V = X * (N * P)$$

Em que:

V = Valor do cálculo

X = Termo da taxonomia encontrada no texto

N = Número inteiro de ocorrências do termo X no documento

P = Peso do termo de acordo com sua posição no texto (Quadro 4)

No princípio acima, o algoritmo computacional embutido no *software* deverá percorrer o documento e identificar, para cada termo da taxonomia: o número de vezes que ele é encontrado no texto, qual a sua localização e realizar o cálculo de V. Realizada essa etapa de submissão do documento ao vocabulário, uma lista de termos encontrados e sua valoração é criada, a fim de realizar uma escolha de quais os descritores apresentarão a melhor representação do documento. O *software* deve ser configurado a fim de permitir a escolha de um número inteiro que permita uma filtragem dos *n* melhores descritores,

seguindo uma ordenação decrescente do sistema de pesos, evitando que pesos muito pequenos participem do conjunto que descreve a obra. Tais descritores estarão registrados em uma estrutura, através do padrão DC de metadados, e gravados no padrão aberto XML. Essa informação deverá ser atrelada ao arquivo da tese/dissertação e será fonte de pesquisa no processo de recuperação da informação.

Concluído o processo de indexação, o ambiente deve ser preparado para a estrutura de recuperação. Este mecanismo será baseado na taxonomia e por palavras e terminologias livres (linguagem natural). A grande vantagem da utilização de um instrumento padronizado na atribuição de termos descritores para os documentos é disponibilizar referências específicas que serão usadas no momento da recuperação.

Desse modo, a interface de recuperação vai contar com um mecanismo hierarquizado para a exibição do vocabulário, permitindo a escolha dos termos em ordem alfabética. A apresentação dos documentos obedecerá à ordem dos pesos atribuídos no processo de indexação. Na existência de um provável empate, a data de criação do arquivo estabelecerá um critério de ordenação decrescente, em que os documentos indexados anteriormente são listados primeiro. Tanto a indexação quanto a pesquisa e o gerenciamento de conteúdo deverão ser feitos dentro do próprio *software* gerenciador.

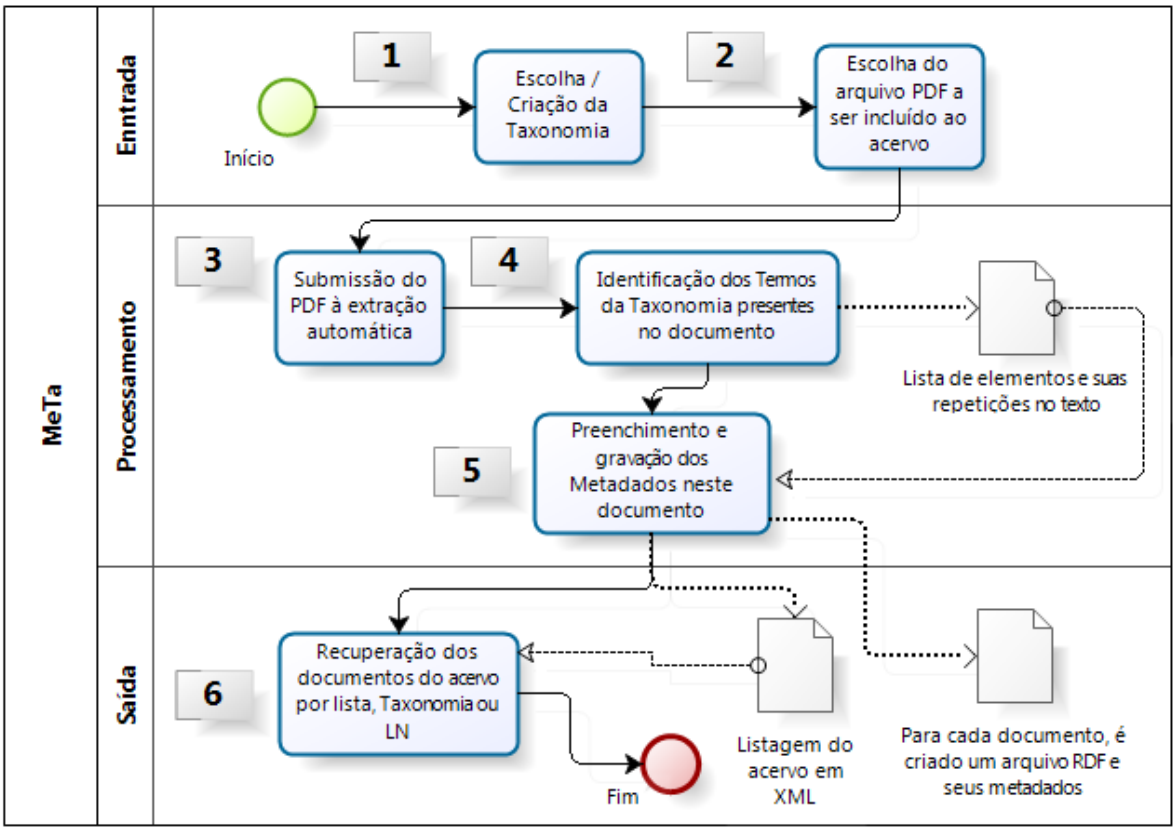
A fim de facilitar o entendimento do usuário, pode-se visualizar um resumo das principais atividades através da sequência apresentada a seguir, organizada em tópicos e figura de processos. Desta forma, o termo “gestor” define um profissional da informação dotado de capacidades técnicas para realizar decisões sobre o vocabulário controlado e outros aspectos técnicos do protótipo. “Usuário” é qualquer pessoa que deseje operar o protótipo a fim de incluir documentos ou recuperá-los.

a) o gestor da informação cria, importa ou seleciona a taxonomia que será a base para a indexação semi automática. Por *default*, o software deve adotar a definição de uma taxonomia inicial. Nesse caso, o *software* assume estes termos como referência para a submissão e a recuperação dos documentos;
b) o gestor/usuário importa o(s) documento(s), tese ou dissertação a ser(em) processado(s);
c) o documento é submetido à taxonomia;
d) o programa gera uma lista de termos encontrados, com base no vocabulário controlado e no sistema de pesos sugerido. O algoritmo realiza os cálculos a fim de selecionar os melhores termos dentro da metodologia escolhida;
e) o programa gera uma estrutura de metadados externa – em arquivo XML/RDF – para cada documento, que será usada na busca de termos do vocabulário;

f) o usuário final pode submeter buscas no documento através da interface do próprio sistema, escolhendo palavras/termos do vocabulário ou da linguagem natural.

O modelo pretende, ainda, permitir que metadados de formatos como MARC, DC e MTD-BR sejam agregados (metadados externos) ao documento, a fim de possibilitar que as informações possam ser exportadas para outros repositórios de Bibliotecas Digitais.

Figura 17 - Macro processos para o protótipo



Fonte: Autor

4 O DESENVOLVIMENTO DO PROTÓTIPO

A partir do embasamento teórico, a construção do *software* teve início através de um modelo de prototipação[36]. A decisão de qual plataforma e qual linguagem utilizar se limitou ao conhecimento prévio e ao custo temporal para aprendizado de outras tecnologias, de modo que a plataforma *Windows*[37] e a linguagem *Object Pascal*[38] foram utilizadas para a implementação do protótipo. Inicialmente foram definidas algumas funções básicas para a construção de uma sequência de atividades que devem resultar na organização e recuperação dos documentos científicos. Partindo da sequência proposta na figura 17, a estrutura constituía-se basicamente de seis grandes grupos:

a) importação/seleção da taxonomia;
b) importação do documento (tese/dissertação);
c) submissão do documento à extração automática de termos baseados na taxonomia;
d) geração de lista de elementos e suas referências no documento baseado na taxonomia;
e) geração de metadados em padrão *Dublin Core* e em outros formatos;
f) seleção do documento mediante o mapa semântico da taxonomia e exibição do enfoque no contexto no documento.

4.1 Estruturas / Interface

Ao se iniciar o protótipo a partir da construção da tela principal, foi definido o modo MDI (*Multiple Document Interface*)[39], que permite a interação com mais de uma janela na mesma interface com o usuário, facilitando sua navegação entre os processos disponíveis no *software*. Esse tipo de interação possui características que potencializam o entendimento do sistema pelo usuário, além de ser usado em uma grande variedade de *softwares* comerciais, amplamente utilizados pela comunidade. A preocupação com a independência de tecnologias para realizar a gestão dos arquivos digitalizados também foi levada em consideração. Para tanto, as configurações básicas, a estrutura da taxonomia e a

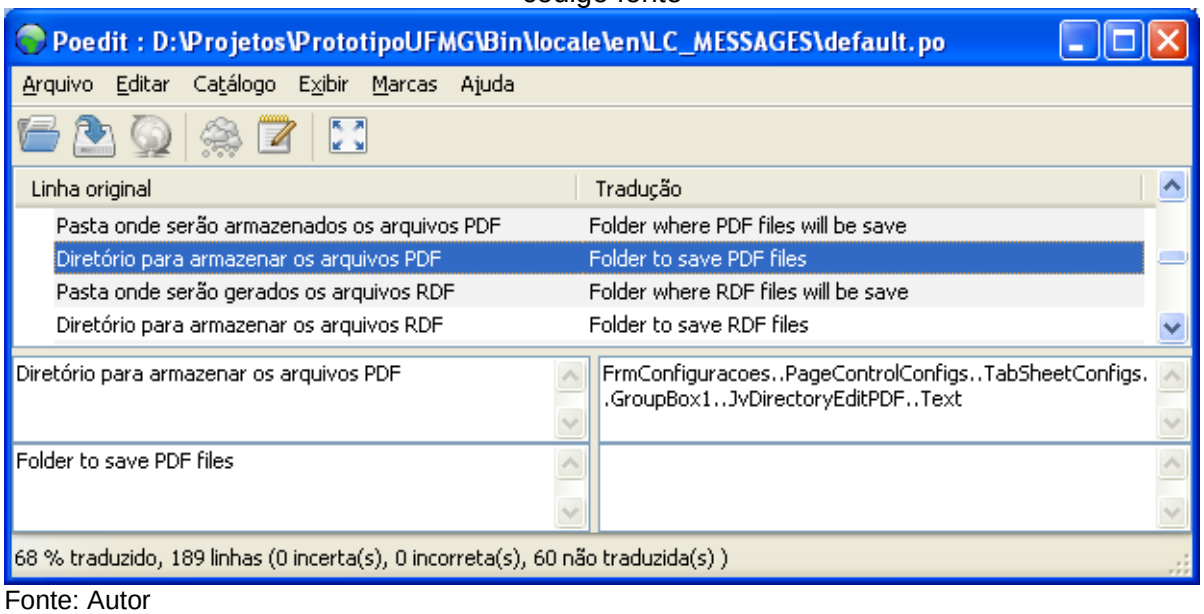
[36] Modelo em que um *software* possui as características funcionais reduzidas do projeto com a finalidade de testar e avaliar suas funcionalidades básicas.
[37] *Windows* é o sistema operacional utilizado nos computadores de plataforma PC (*personal computer*).
[38] *Object Pascal* é uma evolução da linguagem de programação *Pascal*. Em sua versão orientada a objetos, a construção do *software* é baseada nas premissas de classes, objetos e eventos.
[39] É uma técnica de desenvolvimento de *software* em que uma janela principal acondiciona as demais janelas e recursos, criando um ambiente onde as funcionalidades possuem uma dependência visual da janela principal.

estrutura de metadados estão gravadas em formato XML, permitindo sua manipulação por outros sistemas de recuperação de informação.

4.2 Multilinguagem

Com o objetivo de permitir ao usuário escolher o idioma de utilização do software, foi utilizado um *software* para o recurso de localização. O projeto *GNU gettext for Delphi, C++ Builder and Kylix 1.2 beta* é uma metodologia que faz uso de *software* livre e que permite traduzir os textos[40] da interface e caixas de diálogo para diversas linguagens. A tradução é manual e consiste em um processo simples em que o *software* extrai o conteúdo textual passível de tradução do código fonte de programação e disponibiliza uma listagem de frases ou palavras. Através de outro *software* específico (Poedit - figura 18), cada frase é acessada para que a tradução obedeça à mesma referência interna do *software*. O arquivo final é colocado em uma subpasta no projeto do código fonte e, internamente, comandos no código de programação permitem a definição de qual linguagem deve ser usada, permitindo a tradução do *software*. Atualmente, o *software* conta com quatro idiomas: Português, Inglês, Espanhol e Francês.

Figura 18 – Programa que permite traduzir os textos (*Strings*) com sua referência para o código fonte

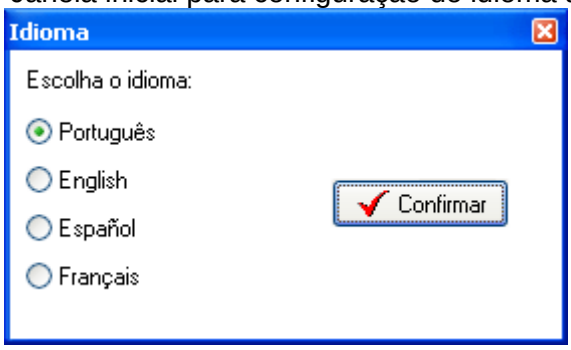


Fonte: Autor

Ao executar o *software* traduzido pela primeira vez, uma janela é exibida, a fim de definir qual o idioma se deve utilizar na interface do sistema. Por padrão, a confirmação define Português (Brasil).

[40] Denominados *Strings* na programação de computadores.

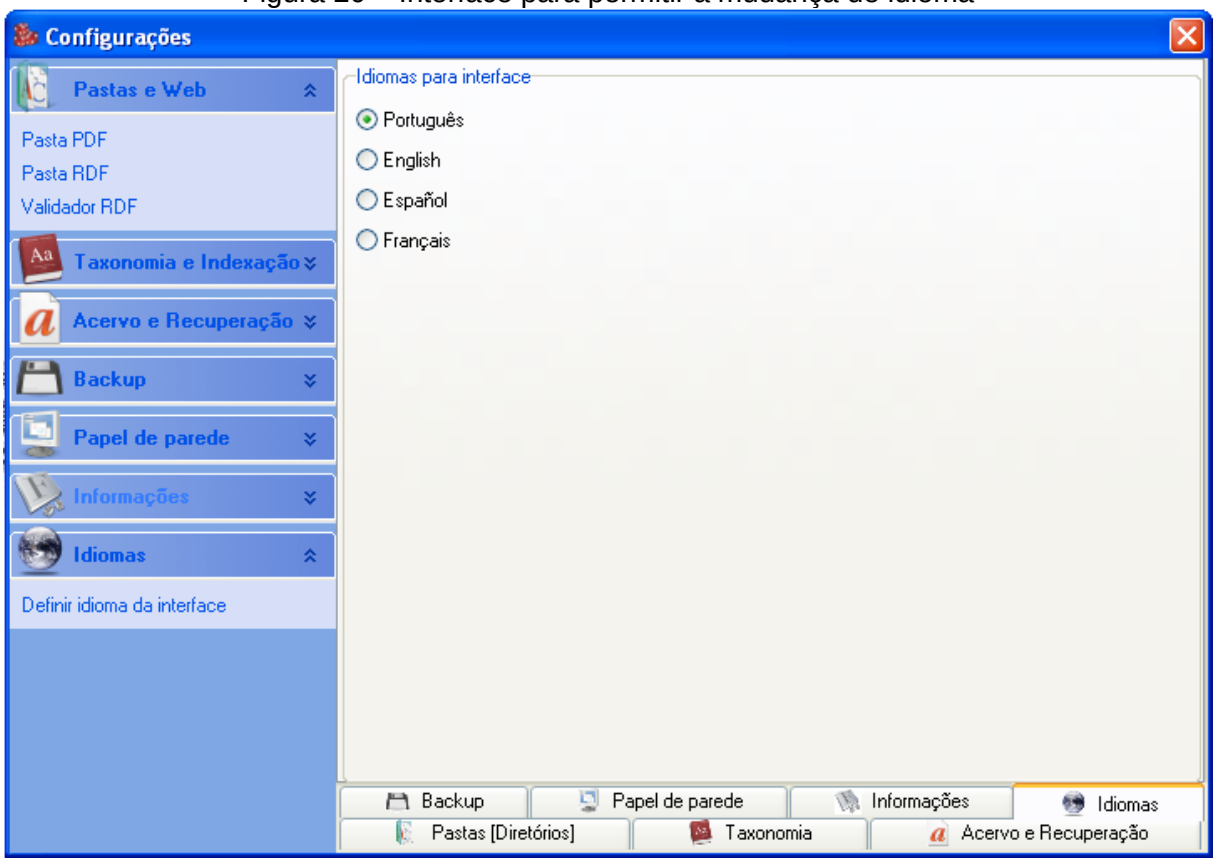
Figura 19 – Janela inicial para configuração do idioma a ser usado



Fonte: Autor

E mesmo após a definição inicial, na janela principal do *software*, através do menu Opções > Configurações (figura 20), é possível definir um outro idioma para a interface.

Figura 20 – Interface para permitir a mudança de idioma



Fonte: Autor

4.3 Configuração

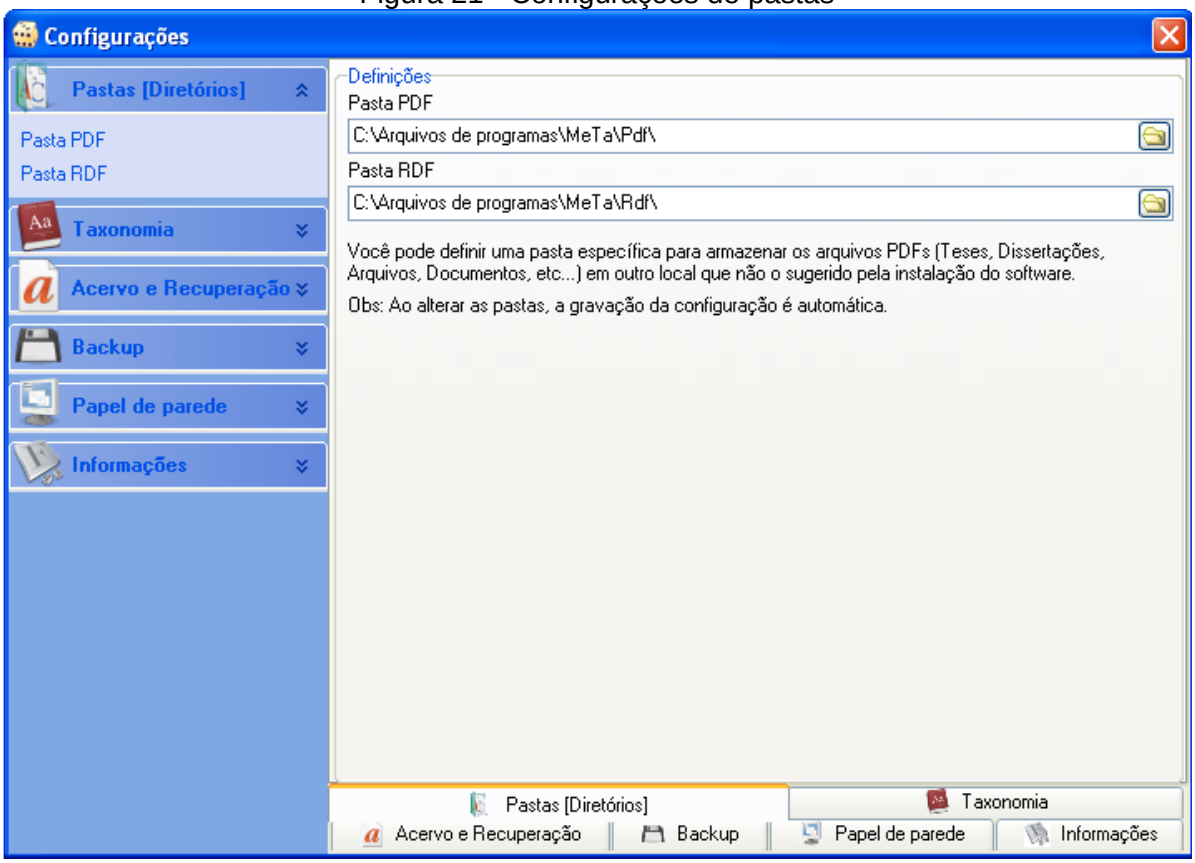
Foi criada uma estrutura para o gerenciamento de diversos itens de configuração do sistema. Essa estrutura pretende definir:

a) a localização física dos arquivos dos documentos em formato PDF (*Portable Document Format*);
b) a localização física da sua representação em arquivos RDF;
c) a indicação de qual taxonomia utilizar;
d) qual arquivo com referência ao acervo (metadados em XML) deve ser considerado pelo *software*;
e) as configurações sobre *Backup*, Papel de Parede, Relatório de variáveis do sistema e qual idioma deve ser usado.

4.4.1 Configurações de Pastas

As configurações de pastas permitem ao usuário final a alteração do local onde os arquivos PDF e RDF serão armazenados. Essa opção é importante para permitir a utilização de unidades de armazenamento externas ao computador local.

Figura 21 - Configurações de pastas



Fonte: Autor

Além disso, esta configuração viabiliza a manutenção e o acesso a outros *softwares* possam ser feitos do modo mais simples e transparente possível.

Os itens de configuração e suas funcionalidades são os seguintes:

Quadro 5 - Itens de configuração

Item	Funcionalidade
Pasta PDF	Nesta pasta (diretório), devem ser gravados os arquivos no formato PDF. O protótipo gravará automaticamente uma cópia do documento nesta pasta, caso ainda não exista. O objetivo principal dessa técnica será explicado no tópico sobre a extração de itens. *Preenchimento obrigatório.*
Pasta RDF	Nesta pasta (diretório), devem ser gravados os arquivos no formato RDF, cujo conteúdo será extraído dos metadados criados no processo de documentação do item digitalizado. O protótipo deve gerar esse arquivo após a gravação dos metadados. *Preenchimento obrigatório.*

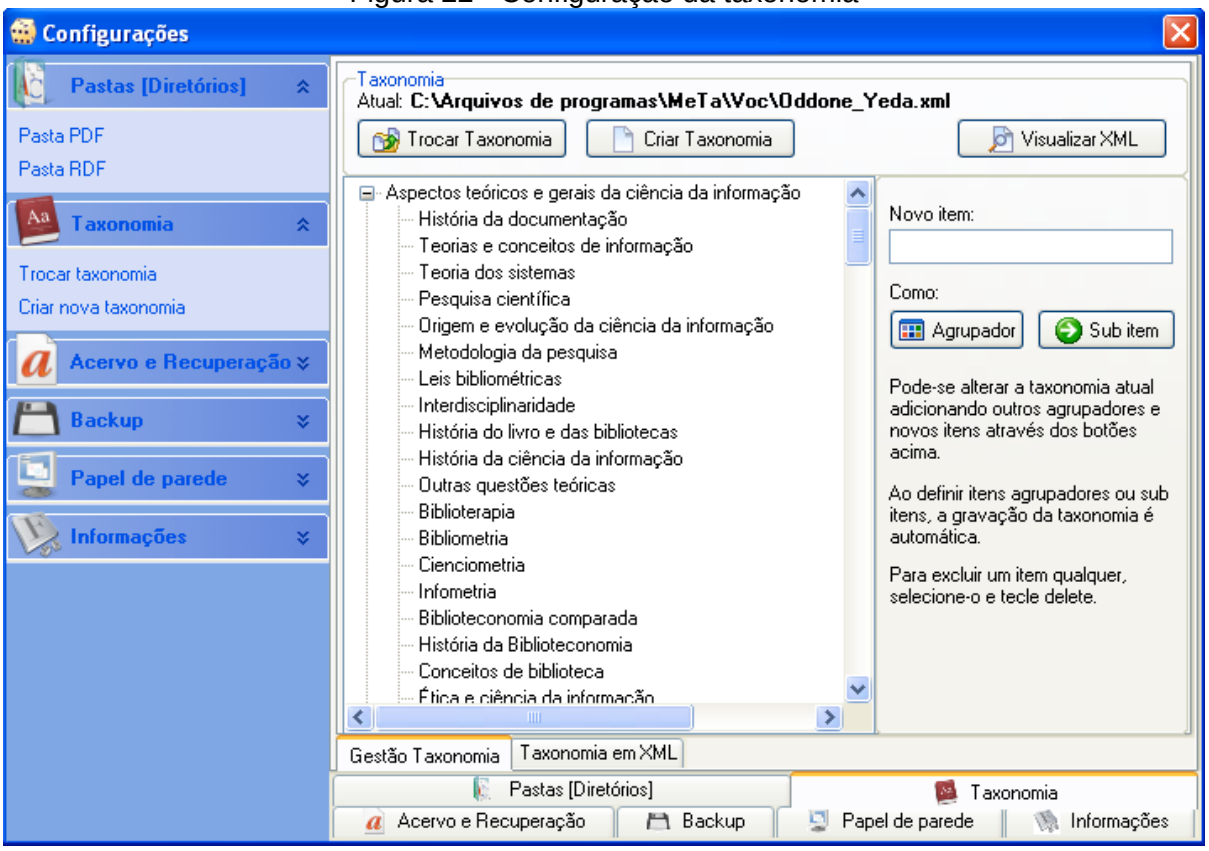
Fonte: Autor

Quando o *software* é instalado, esses itens são pré-configurados, permitindo que o usuário os altere, personalizando-os.

4.4.2 Definições / Seleção da taxonomia

Outro item passível de configuração, a Taxonomia será responsável por direcionar tanto a indexação como a recuperação dos itens registrados no protótipo. Sua adoção é obrigatória, e toda a estrutura do sistema é baseada nos termos cadastrados. A intenção é que o *software* possa trabalhar com a escolha de taxonomias personalizadas, não limitando o procedimento a uma lista fixa.

Figura 22 - Configuração da taxonomia



Fonte: Autor

Esse instrumento inicialmente foi idealizado através de uma lista sequencial de termos que seriam usados para estabelecer uma estrutura de classificação do item.

Desse modo, foi considerada a adoção da estrutura terminológica baseada na taxonomia do artigo de Oddone e Gomes (2004), que ainda faz referência ao trabalho de Hawkins; Larson; Caton (2003). Assim, a Taxonomia foi organizada em dez categorias e suas respectivas subcategorias.

Essa taxonomia é definida como padrão no momento da instalação do *software*, mas existe a possibilidade de alteração dessas configurações através de mecanismos de inclusão e exclusão de itens, além de ser possível a criação de outra estrutura Taxonômica de modo totalmente independente. Essa flexibilidade é importante para permitir que o programa seja utilizado por diferentes áreas do conhecimento.

Para abordar outras possibilidades deste instrumento, o desenvolvimento de algoritmos para importar taxonomias gravadas em outros formatos também será necessário, porque pode-se encontrar termos em outros formatos, tais como TXT[41], XLS[42], entre outros.

[41] Formato de texto puro, sem formatação, que pode ser lido e alterado por diversas plataformas de *software*.
[42] Formato de planilha eletrônica, muitas vezes utilizada para a organização de listas.

Atualmente, os dados são gravados em arquivos XML e trazem uma correspondência interna para permitir que as categorias agrupem as subcategorias.

Tecnicamente, o mecanismo construído no protótipo era muito próximo do projeto inicial e constituía a base do projeto, pois tanto o mecanismo de inserção do documento no acervo quanto a recuperação dependiam do instrumento a ser usado como taxonomia. Os termos seriam escolhidos, em um primeiro momento, por sugestão do algoritmo, relacionando os termos encontrados no documento em partes específicas, mas o desenvolvimento do sistema de pesos para valoração do termo encontrou diversas dificuldades, de modo que foi adotado um mecanismo de análise de repetição simples.

4.5 Importação do documento

Após a definição das configurações iniciais (pastas, idioma) e taxonomia a ser utilizada, o protótipo está apto para ser manipulado na importação do documento. Os demais passos para a classificação do documento e sua consequente organização através dos metadados são baseados na caracterização do processo de indexação que pode ser visto em seguida.

Foram identificados os seguintes desafios:

a) os documentos estão em formato proprietário?
b) os documentos são protegidos contra edição?
c) os documentos são baseados em texto ou são escaneados (baseados em imagem)?
d) os documentos podem estar em redes externas[43] ao ambiente (computador ou rede) do protótipo?
e) os documentos possuem *copyright*?

Para cada questão acima, houve uma abordagem específica a fim de realizar a documentação. As abordagens são descritas a seguir.

4.5.1 Documentos em formato proprietário

No princípio do projeto, já era sabido que os documentos científicos eram – em sua maioria – publicados no formato PDF[44]. A adoção desse formato deve-se ao fato de sua proposta envolver os seguintes aspectos:

[43] Ambiente computacional da instituição de origem ou Internet.
[44] Sigla para *Portable Document Format*, em português, Formato de Documento Portátil.

a) Portabilidade: esse formato permite representar a informação independente do Sistema Operacional e *Hardware*, sendo que qualquer ambiente (*desktops*, *notebooks*, dispositivos móveis) pode possuir um leitor codificado para esse formato, pois é um padrão aberto;
b) Flexibilidade: pode conter informação textual, gráficos e imagens;
c) Segurança: o arquivo pode evitar a edição e cópia parcial do conteúdo sem autorização.

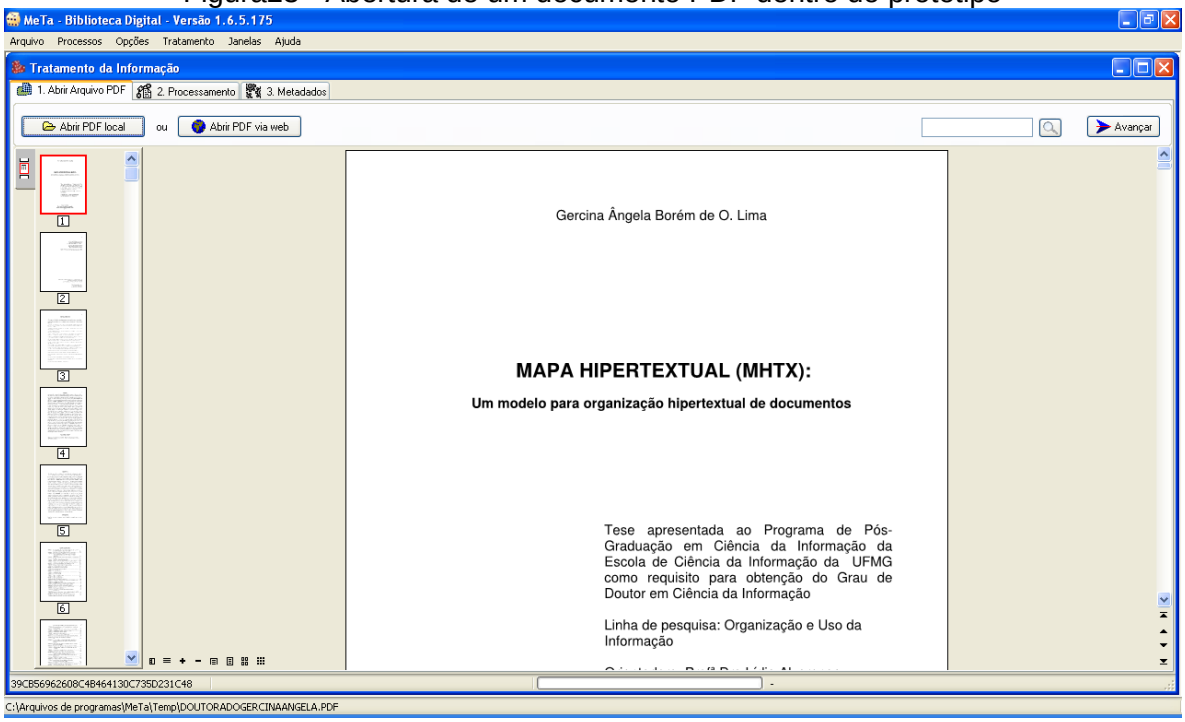
Sendo a estratégia de publicar documentos nesse formato uma característica não somente da comunidade acadêmica, mas dos usuários em suas diversas categorias, a manipulação direta do conteúdo do arquivo em PDF oferecia certo entrave na sequência de passos do processo de classificação, visto que grande parte dos documentos apresentava configurações em níveis de proteção, impedindo cópias parciais e até mesmo sua impressão.

Para resolver a situação de visualização e consequente leitura, foi encontrada, na Internet, uma biblioteca de código compatível com a linguagem de programação adotada no protótipo e de uso gratuito: *Synactis PDF Viewer*[45]. Esse componente[46] possibilita o acesso ao arquivo PDF de modo nativo, sem a necessidade da abertura de um programa externamente (figura 23).

Tecnicamente, o programa padrão para realizar o acesso de leitura ao documento PDF é o *Acrobat Reader* da empresa Adobe. Esse *software* deve estar instalado no computador em que o protótipo realiza seus processos, porém, através da nova codificação, o *Synatics PDF Viewer*, o protótipo utiliza os recursos nativos do *Acrobat* dentro do próprio ambiente, sendo necessária apenas a instalação do *software* no computador em que o protótipo está instalado.

[45] http://www.synactis.com/
[46] Código computacional que realiza um processo específico.

Figura23 - Abertura de um documento PDF dentro do protótipo



Fonte: Autor

Este mecanismo permite que o protótipo trabalhe de forma mais independente, minimizando a desorientação do usuário causada pela manipulação de mais de um programa, em interfaces diferentes.

Esse mecanismo solucionou de modo aceitável o problema de leitura do padrão de arquivo proprietário, mas ainda havia o problema da extração das informações do documento.

4.5.2 Documentos baseados em texto ou imagens

Quando os documentos são baseados em textos, pode-se observar as situações citadas anteriormente quanto à leitura e extração dos caracteres. Porém, quando o documento é baseado em imagens, total ou parcialmente, elas não são recuperadas na forma textual, formato presente na interface de recuperação.

A complexidade desse processo remonta aos *softwares* de reconhecimento OCR[47]. O protótipo não vai abordar a inclusão dessa tecnologia no processo, de modo que as imagens não serão consideradas nos algoritmos de sugestão dos termos e pesquisa/recuperação. Mesmo assim, um documento baseado em imagens poderá ser indexado através de atribuição de termos de forma manual pelo usuário.

[47] *Optical Character Recognition*, que pode ser traduzido como Reconhecimento Ótico de Caracteres.

4.5.3 Documentos que possuem *copyright*

O protótipo partiu do princípio de que todo documento científico já possui direitos autorais. Desse modo, embora o arquivo precise ser manipulado e tenha seu conteúdo extraído externamente para o PDF, ele não é disponibilizado para a edição do usuário ou mesmo cópia em memória. Após o processamento interno, o que permanece na organização física dos arquivos do acervo é o documento original, em seu formato PDF.

4.6 Identificador

Outro detalhe importante foi a geração de um código único para o documento, a fim de identificá-lo e permitir validações futuras, além de planejar uma referência única para os metadados.

Para isso, o arquivo PDF é submetido a um algoritmo denominado MD5[48]. Esse código, ao receber um texto (*String*), retorna uma codificação textual que representa o identificador do arquivo independente de seu nome físico, pois o código "lê" o conteúdo do arquivo PDF.

Esse código, exemplo: "39CB56962608C4B464130C735D231C48", vai permitir que arquivos com nomes diferentes (fulano.pdf e beltrano.pdf), embora com o mesmo conteúdo, sejam avaliados antes das etapas de entrada e indexação no acervo digital, evitando duplicidades e a geração de metadados diferentes pretendendo descrever o mesmo arquivo.

Adiante aborda-se a aplicação desse código no preenchimento dos metadados.

4.7 Extração automática de termos baseados na taxonomia

Até este momento, o fluxo do processo seguiu a seguinte sequência no processamento do documento:

a) definição das configurações iniciais: taxonomia, pastas, idioma;
b) abertura do arquivo PDF para visualização e geração do código identificador.

É, então, tempo de avançar para processar o arquivo frente à taxonomia. Esse momento consiste em confrontar o texto com os termos presentes no instrumento definido. Um algoritmo de contagem para as repetições dos termos encontrados no documento pretende sugerir ao indexador (humano) quais os melhores descritores para o documento,

[48] *Message-Digest algorithm 5* – desenvolvido pela *RSA Data Security* http://www.ietf.org/rfc/rfc1321.txt

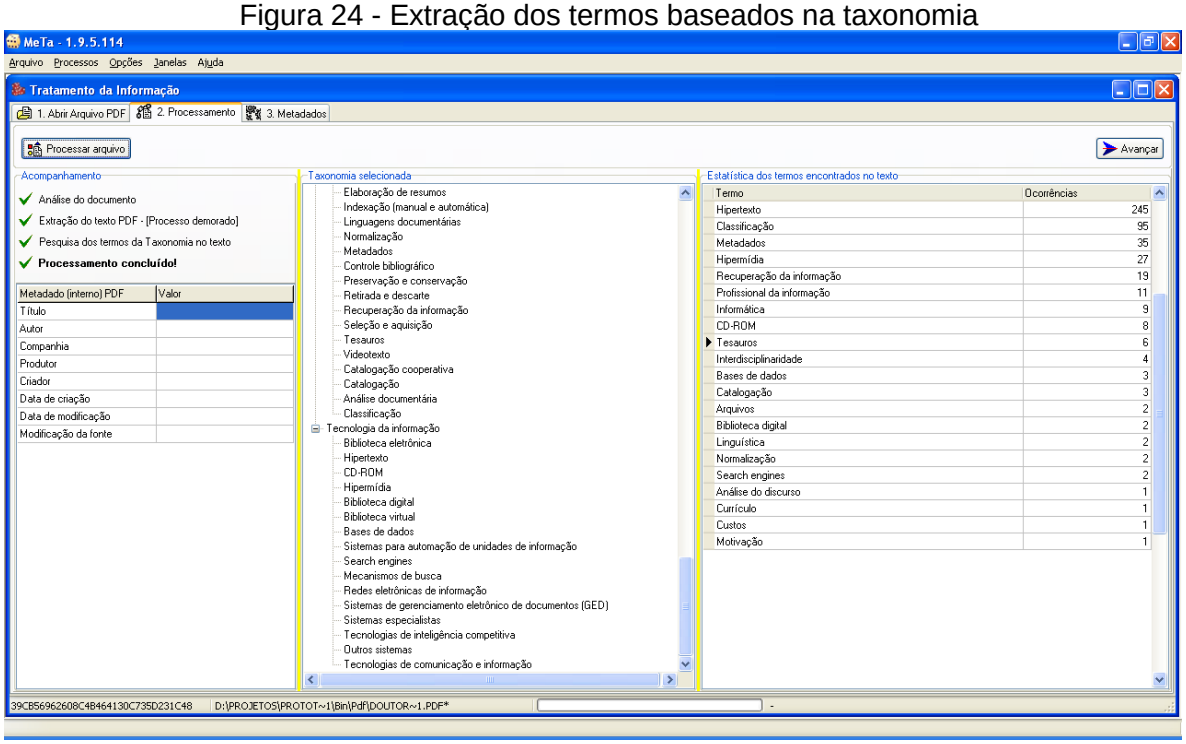
deixando em aberto a condição de inserção de outros termos externos ao vocabulário, de maneira manual.

Nesta etapa, foi considerado utilizar um *software* externo para extrair o texto do documento PDF, uma vez que esse formato não permite um processamento direto em seu conteúdo. Os problemas dessa abordagem poderiam ser elencados da seguinte forma:

a) para utilizar o protótipo, o usuário precisaria instalar mais um *software*;
b) essa etapa do processamento seria interrompida no protótipo para abertura e manipulação de outro *software* que pode exigir manuseio e configurações particulares;
c) não haveria um controle do protótipo sobre o *software* externo, dificultando a sequência de etapas a serem seguidas.

Para trabalhar essa condição limitadora, utilizou-se outra biblioteca de funções de *software*, desta vez através de um código comercial sem necessidade de *royalties*[49]. A solução *PDFtext* (INGO SCHMOEKEL, 2012) permite ao desenvolvedor extrair o texto do PDF e retornar ao fluxo de processos desejado, de modo que o texto extraído pode ser trabalhado de forma livre, sem restrições e proteções dentro do código fonte do protótipo.

Figura 24 - Extração dos termos baseados na taxonomia



Fonte: Autor

[49] *Royalty* é uma palavra de origem inglesa que se refere a uma importância cobrada pelo proprietário de uma patente de produto, processo de produção, marca, entre outros, ou pelo autor de uma obra, para permitir seu uso ou comercialização. (BRASIL, 2012c).

Essa possibilidade de extração, porém, não pretende abrir o documento de maneira permanente ou desautorizada. A extração de textos para processamento é uma função interna ao sistema, sem a intervenção humana. E, caso o documento possua uma proteção mais específica como a adoção de senhas, o algoritmo retorna uma mensagem de impossibilidade de realização do processo. Mesmo assim, a indexação poderá ser realizada manualmente com a adoção de termos a serem definidos pelo usuário.

Os principais tópicos a serem analisados na figura 25 são:

a) caso o arquivo PDF possua metadados no momento de sua geração, eles serão exibidos nesse momento e usados posteriormente como sugestão no cadastro do item no acervo do protótipo;
b) os termos do vocabulário por categoria são os termos confrontados com o texto puro [1], a fim de estabelecer uma estatística;
c) a estatística dos termos do vocabulário encontrados no texto é exibida em ordem de repetição decrescente.

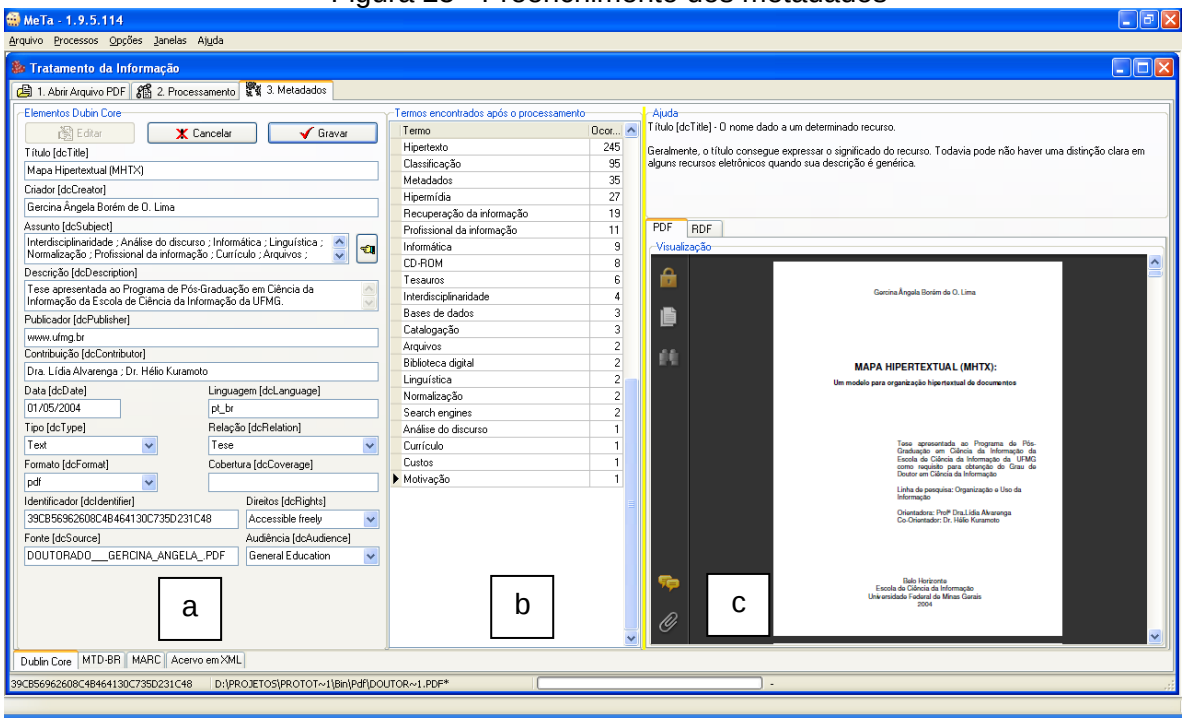
Neste ponto, o processo não define os metadados, foco deste trabalho, possui apenas uma interface de observação para um acompanhamento didático do processo. Na próxima etapa, os termos[50] mais "pesados"[51] encontrados nesta fase serão sugeridos como possíveis descritores do documento, para que sejam usados no elemento Descrição (*dcDescription*).

4.8 Gravação de metadados

Após o processamento do texto, que consistiu em extrair o texto em formato puro do PDF e confrontar os termos da taxonomia, buscando uma estatística das repetições terminológicas, o protótipo permite a gravação dos metadados, inicialmente no padrão *Dublin Core*.

[50] Neste trabalho, a terminologia Elementos e Campos possuem o mesmo significado.
[51] Entendem-se como termos "pesados" os de maior ocorrência no texto.

Figura 25 - Preenchimento dos metadados



Fonte: Autor

A interface que disponibiliza o processo, como visto na figura 25, é composta pelos seguintes itens na guia *Dublin Core*:

a) elementos a serem preenchidos manualmente em sua maioria. Alguns campos possuem um valor de sugestão, a ser apresentado adiante;
b) termos selecionados pelo algoritmo que confrontou o texto puro com o vocabulário controlado, em ordem decrescente estatística. Esse conjunto pretende sugerir valores, os quais poderão ser utilizados na recuperação, usando o campo *Subject*.
c) visualização do documento original, formato nativo, PDF como fonte de fácil acesso para o preenchimento dos elementos. Ainda nessa parte da interface, após a gravação dos metadados, é possível visualizar o código dos metadados em RDF, criado para descrever o documento após a gravação.

Os campos para os metadados: *type*, *format*, *identifier*, *source*, *rights* e *audience* são preenchidos com valores *default*, após o processamento do texto na etapa anterior. A lógica de preenchimento para cada campo foi a seguinte:

a) *Type* – como o trabalho tem como objeto principal arquivos, em sua maioria, textuais, o valor *Text* foi atribuído a esse campo. As outras opções,

recomendações da DCMI[52], estão disponíveis pelo recurso caixa de listagem presente no campo do elemento;

b) *Format* – nesse campo, foi sugerido o valor *PDF*, que descreve o formato do arquivo original;

c) *Identifier* – a DCMI define o valor desse campo como um *identificador único*, que pode especificar o recurso em determinado contexto. O documento PDF, ao ser aberto, como descrito anteriormente, é submetido ao algoritmo MD5, que retorna um código único que permite identificar esse documento;

d) *Source* – esse elemento é preenchido com o *nome físico do documento*. Uma vez que nas configurações, pode-se ter acesso à pasta onde estão armazenados os arquivos PDF, é possível localizá-los através da junção dessas duas informações;

e) *Language* – é preenchido com o valor *pt_br*, por padrão, pois o protótipo está sendo lançado inicialmente na comunidade acadêmica brasileira;

f) *Rights* – campo preenchido com *Accessible freely,* significando o livre acesso aos documentos;

g) *Audience* – é preenchido com o valor *General Education* que pretende focar o público acadêmico.

Todos os valores citados acima podem ser alterados manualmente pelo usuário. Os valores *default* são sugestões que pretendem agilizar o preenchimento dos metadados no padrão *Dublin Core*.

Após a gravação das informações, um arquivo no formato XML é atualizado com as informações, com vistas à recuperação futura, e o processo pode ser realizado para outros documentos.

4.9 Recuperação da informação no protótipo

O processo de recuperação no protótipo permite uma abordagem em três grupos principais, que podem utilizar:

a) Listagem Completa
b) Taxonomia
c) Linguagem Natural

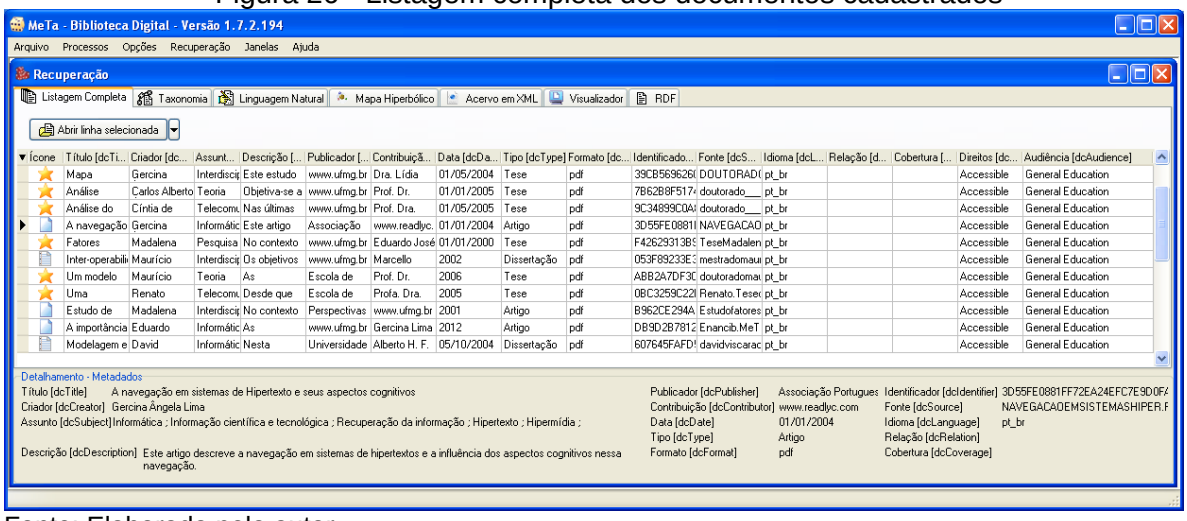
[52] *Dublin Core Metadata Initiative*

Todo esse processo depende da fase anterior, em que a inclusão do documento no acervo permitiu definir os metadados e outras informações que pretendem atuar como termos para recuperação do item eletrônico.

4.9.1 Pesquisa através da Listagem Completa dos itens

No primeiro momento, a interface de pesquisa exibe uma listagem dos itens cadastrados no protótipo, permitindo, através da ordenação de colunas, localizar um documento pelos metadados em *Dublin Core*.

Figura 26 - Listagem completa dos documentos cadastrados

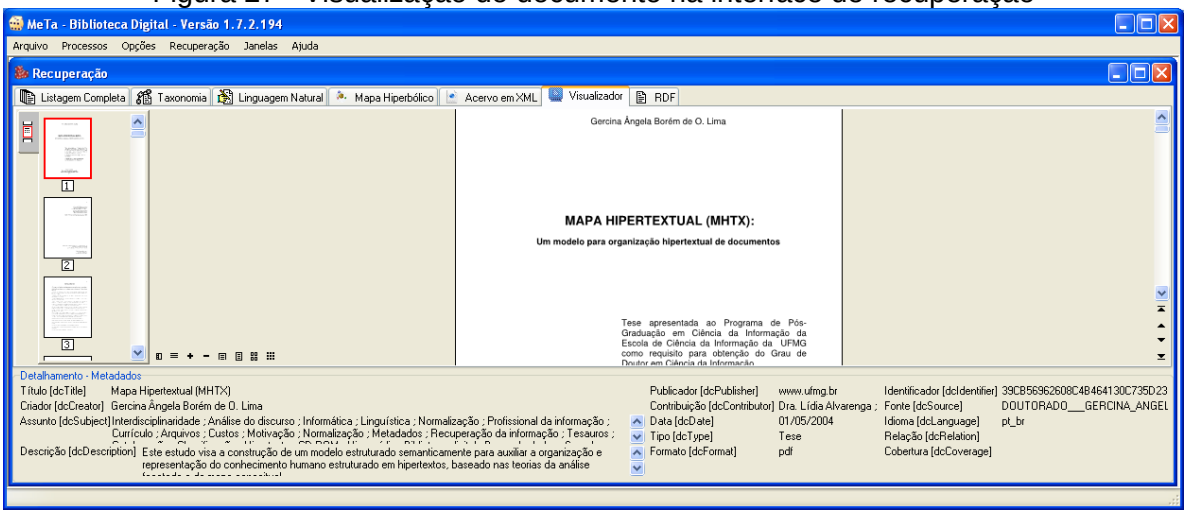


Fonte: Elaborado pelo autor

Caso o usuário identifique o documento e deseje sua visualização, pode clicar duas vezes rapidamente na linha para que o protótipo abra[53] o PDF na própria interface, através da guia *Visualizador,* como na figura 27.

[53] A abertura do arquivo para leitura pode ser configurada de modo a usar um leitor externo ao protótipo. Essa opção está descrita no tópico de configurações.

Figura 27 - Visualização do documento na interface de recuperação



Fonte: Elaborado pelo autor

Esse tipo de pesquisa pretende facilitar situações nas quais a ordem alfabética pelo título ou autor, ou outro metadado que possa ser ordenado, permita a localização de um item dentro do acervo.

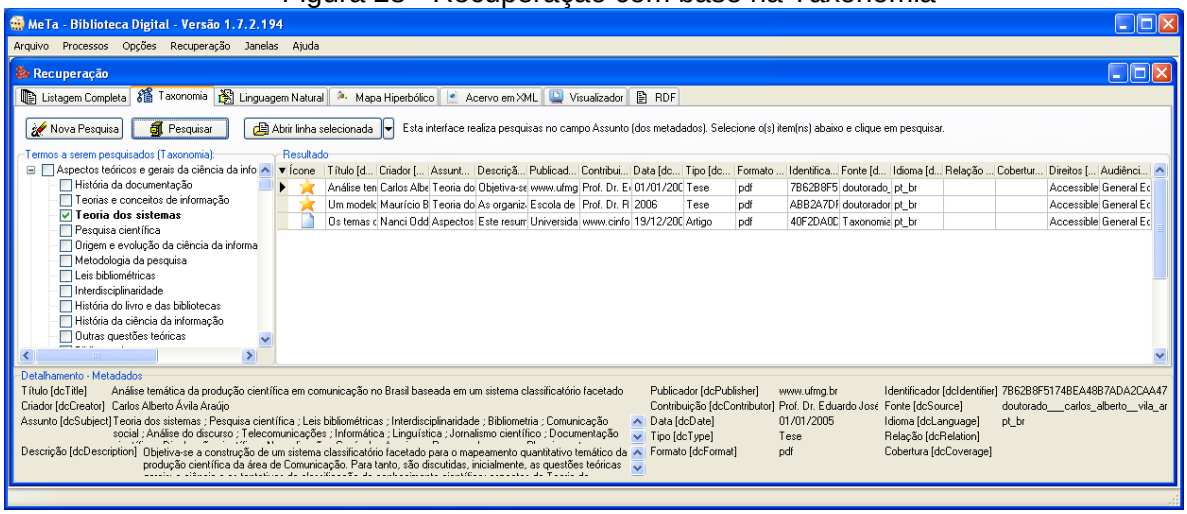
4.9.2 Pesquisa com base na Taxonomia

Outra possibilidade é recuperar o(s) documento(s) através da taxonomia, instrumento usado no processo de inclusão do item no acervo. Ou seja, os termos eleitos como elementos que representam o documento em potencial foram incluídos no metadado assunto (*dcSubject*) no momento do cadastro e podem ser usados nessa interface para realizar a pesquisa e recuperação dos itens.

No exemplo da Figura 28, foi selecionado o termo “Teoria dos Sistemas”. A pesquisa retornou 3 (três) documentos que possuem essa terminologia em um campo específico dos seus metadados.

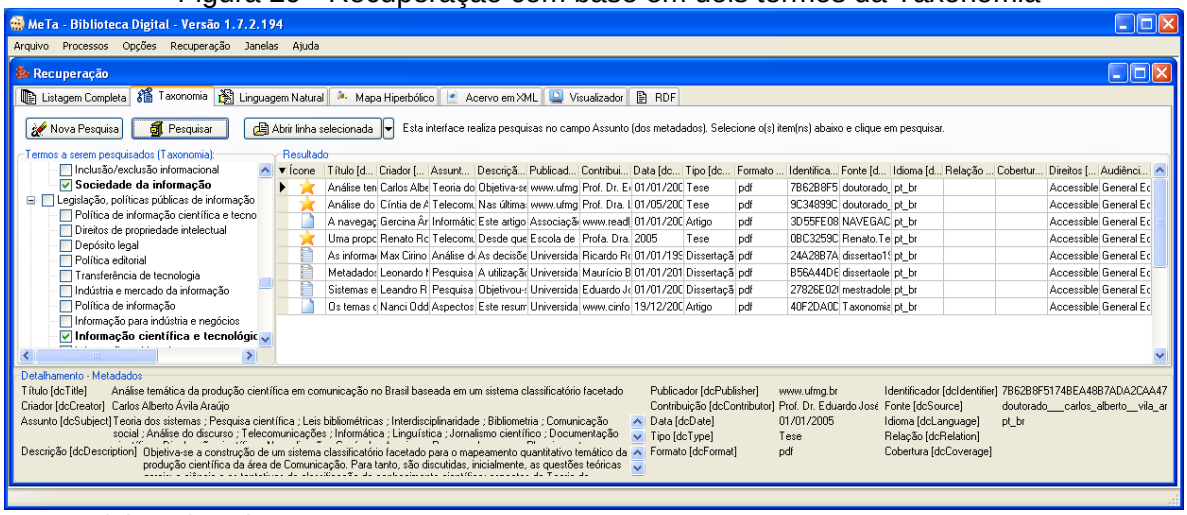
Também é possível considerar vários termos para serem pesquisados ao mesmo tempo. Ao selecionar os termos “Sociedade da informação” e “Informação científica e tecnológica”, a pesquisa anterior, que havia retornado 3 documentos, agora retorna 8 (oito) documentos (figura 29), devido à inclusão de mais um termo como critério de pesquisa.

Figura 28 - Recuperação com base na Taxonomia



Fonte: Elaborado pelo autor

Figura 29 - Recuperação com base em dois termos da Taxonomia



Fonte: Elaborado pelo autor

A importância da taxonomia utilizada na indexação é evidente na capacidade de fornecer termos e, principalmente, ajudar na recuperação dos itens do acervo eletrônico.

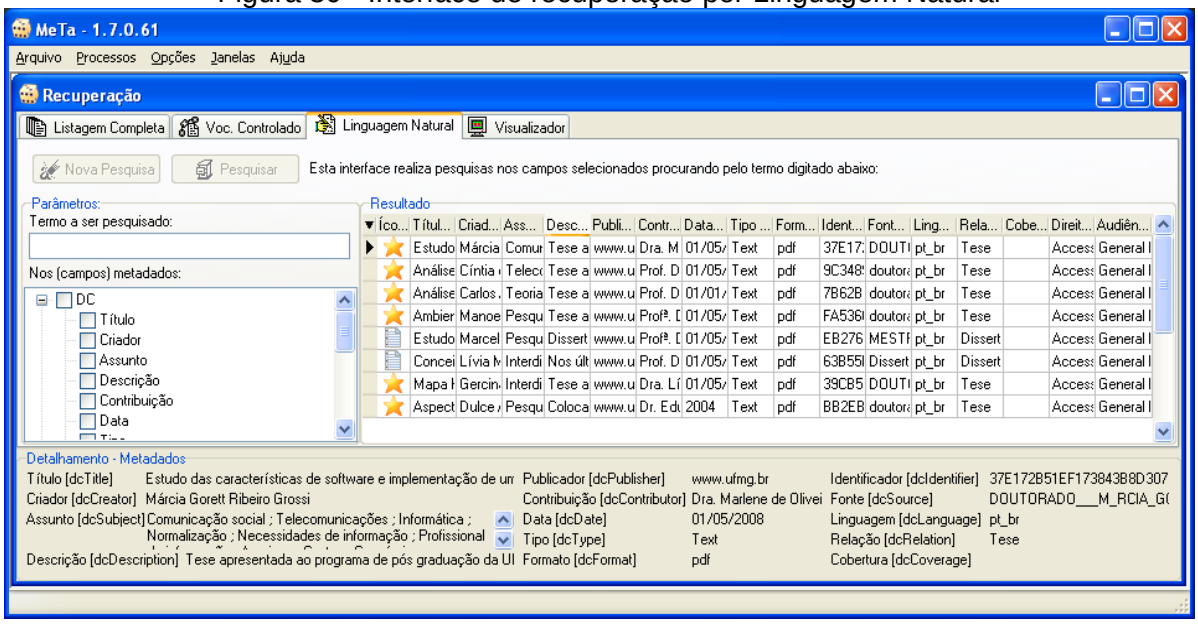
Nesse protótipo, os termos eleitos para representação foram cadastrados no metadado Assunto (*dcSubject*) de modo que, para haver um mecanismo que permitisse a pesquisa nos outros campos, era necessário possibilitar o acesso às informações de outra forma, na qual a taxonomia não foi usada. Esse mecanismo foi criado através da Linguagem Natural.

4.9.3 Pesquisa através da Linguagem Natural

A pesquisa em linguagem natural, livre em sua construção, permite, no protótipo, o acesso aos demais campos e a suas informações.

Ao definir o termo a ser pesquisado, a interface do protótipo permite a escolha de pesquisa em um ou mais campos específicos através das caixas de seleção, como mostra a Figura 30 abaixo.

Figura 30 - Interface de recuperação por Linguagem Natural

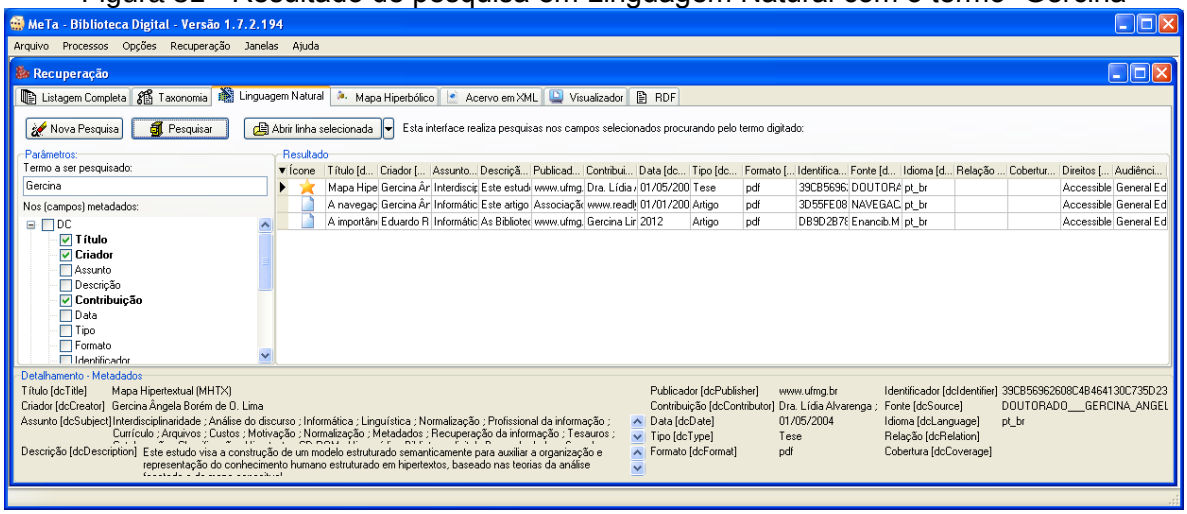


Fonte: Elaborado pelo autor

O direcionamento da pesquisa ocorre em função dos campos da linguagem documentária disponível, no caso, o *Dublin Core*. Diferentemente da recuperação pela taxonomia, que seleciona os termos em função de sua associação ao item bibliográfico, a linguagem natural é mais livre e, ao mesmo tempo, mais desafiadora, pois permite ao usuário requisitar termos que não foram usados na catalogação/indexação e exprimem um sentido sinônimo ou semântico à recuperação.

Na interface do protótipo, o usuário pode recuperar a informação direcionada ao tipo de informação previamente cadastrada. Por exemplo, ao procurar o termo “Gercina”, selecionando o campo *Título*, nenhuma resposta será exibida. Ao selecionar o campo *Criador*, uma resposta será disponibilizada pelo protótipo, pois, no processo de catalogação, esse termo foi inserido como Criador da obra. E, ainda, ao selecionar o campo Contribuição, dois registros são retornados, pois o termo Gercina foi inserido em uma obra como Criadora e em outra obra como Contribuidora (figura 31).

Figura 31 - Resultado de pesquisa em Linguagem Natural com o termo “Gercina”



Fonte: Elaborado pelo autor

Esta abordagem permite ainda o acesso aos demais campos da linguagem documentária, ressaltando a necessidade de deixar claro ao usuário que o sucesso da recuperação da pesquisa está diretamente ligado às suas escolhas na definição dos termos a serem procurados.

O mecanismo de busca foi construído para utilizar prefixos ou sufixos da terminologia utilizada, de modo que, para encontrar o nome próprio “Gercina”, não foi preciso completar prenome e sobrenome, pois o sistema localiza todo o conteúdo que contém esse termo naquele campo selecionado. O mesmo raciocínio seria válido se fosse digitado o último nome da pessoa à qual a pesquisa se refere: Lima.

5 CONSIDERAÇÕES E RECOMENDAÇÕES

O crescente e visível aumento de publicações tem demandado novos mecanismos de gestão e organização da informação. Nesse sentido, este projeto procurou contribuir com os sistemas de recuperação da informação com enfoque em Metadados.

Além do papel de armazenar os termos, os metadados atuam ainda como tradutores. Embora este projeto se concentre em documentos científicos (teses e dissertações) e em um formato de arquivos pré-determinados (PDF), ao cumprir o mesmo objetivo de catalogação e indexação em arquivos com informações em áudio, imagens ou vídeo, em que a simbologia textual não se faz presente, os metadados assumem um papel ainda mais importante. Nesse caso, eles intermediam linguagens diferentes, permitindo ao usuário a recuperação da informação através da linguagem textual. Na maioria dos sistemas de pesquisa, esse tipo de linguagem funciona como interface de comunicação entre o processo indexador e a recuperação.

Neste trabalho, os metadados surgiram como ponto de atenção em razão da necessidade de se trabalhar em informações que, em sua natureza, não são estruturadas. A inclusão dos metadados no projeto permitiu estruturar a informação não estruturada, o que, para os mecanismos de busca, é uma estratégia de grande valor. Essa abordagem foi fundamental na construção do protótipo, pois as estruturas de dados que trabalharam a gravação e recuperação são altamente estruturadas. Os algoritmos de pesquisa precisavam de estruturas bem definidas para permitir a recuperação, e o próprio mecanismo de armazenamento permanente (não volátil) usado – arquivos em formato XML – exige uma estruturação. Assim, com a linguagem documentária alinhada aos metadados, foi possível criar um mecanismo eficiente de armazenamento e recuperação das informações, seja para a descrição do item bibliográfico (catalogação), seja para os termos de assunto (indexação).

No que se refere à implementação do protótipo, um ponto importante diz respeito à plataforma de implantação tecnológica, uma vez que o protótipo MeTa não foi criado para o ambiente *web* no momento desta pesquisa. Consequentemente, sua utilização se efetiva no modelo *desktop* de funcionamento local ou em ambiente LAN[54]. A codificação para que o *software* seja implementado no ambiente *web* é um projeto futuro.

Outra observação pertinente é a diferença do modelo de arquivos observado no protótipo Mapa Hipertextual (MHTX), que utiliza o formato HTML. O protótipo MeTa foi construído visando, inicialmente, o formato PDF na definição do acervo e para a recuperação dos documentos. Esse padrão é considerado, no momento da conclusão desta pesquisa, o mais utilizado para a publicação das obras. Contudo, como esse formato não é viável de referência direta interna (a fim de identificar páginas e parágrafos específicos por

[54] *Local Area Network* – Rede de abrangência local.

algoritmos de terceiros), também não foi possível realizar a recuperação no contexto específico do documento, projeto a ser implementado futuramente. Dessa maneira, no protótipo atual, os termos para a recuperação e a linguagem natural sempre identificam o arquivo como obra completa e, não, como parte (parágrafo/página) específica do documento.

Durante o desenvolvimento do protótipo, notou-se que os metadados associados ao documento trouxeram outras possibilidades, como a de trabalhar com outros arquivos de formatos diferentes do PDF. Os processos foram desenvolvidos com uma visão em documentos textuais e para um tipo específico. A visualização e extração foram codificadas de modo direcionado ao formato PDF, mas a aplicação do modelo independe dessa característica textual do documento a ser trabalhado, o que permite a indexação e recuperação de arquivos multimídia e em outros formatos, no futuro, sem grandes alterações no código computacional.

A respeito do padrão de metadados adotado, o protótipo pretende também trabalhar com outras linguagens documentárias em estudos futuros, como o MTD-BR e MARC21, como alternativa ao *Dublin Core*, com vistas a ampliar as opções de recuperação da informação e permitir a exportação (compatibilização) para outros acervos, porém, somente a implantação em *Dublin Core* foi realizada a tempo na presente versão. Essa iniciativa é um desejo de continuidade do projeto também expressa no protótipo nas guias MTD-BR e Marc21.

O processo de recuperação da informação é a parte que mais necessita de alterações no futuro, a fim de se aproximar do modelo MHTX de navegação contextual em sua essência. Na versão atual, foi adotada a recuperação por linguagem natural e por linguagem documentária, que permitem ao usuário a recuperação de um ou mais itens dentro de uma seleção de um determinado termo da Taxonomia, ou pela livre digitação de um termo desejado. Essa etapa foi, sem dúvida, a mais desafiadora do projeto.

O quadro 6 pretende demonstrar alguns dos principais pontos pretendidos, alcançados e não alcançados no protótipo, tendo como base o parâmetro do modelo MHTX.

Quadro 6 - Processos principais

Item / Processo	Objetivo	MeTa	MHTX
Importação/Seleção da linguagem documentária.	Definir um mecanismo para permitir a indexação dos documentos.	Aplicado	Não aplicável
		Através de Taxonomia, o vocabulário controlado é um instrumento importante na indexação e consequente	

		recuperação do item no acervo.	
Importação da Tese/Dissertação	Permitir a seleção do documento para análise e inclusão no acervo.	Foi definido o padrão PDF para a construção do acervo. Um ponto positivo é a larga utilização desse formato. O ponto negativo é, neste momento, a impossibilidade de referenciar uma parte específica (página, parágrafo) dentro do arquivo.	No sistema Greenstone foi usado XML na criação do índice invertido e no software StarTree foi definido o padrão HTML para possibilitar a navegação em contexto, visto que esse formato permite a marcação através de *tags* (âncoras), identificando uma parte específica do texto.
Submissão da tese/dissertação à extração semiautomática baseada na linguagem documentária.	Permitir confrontar o texto com a linguagem documentária (taxonomia), a fim de selecionar os melhores descritores do documento.	Aplicado	Não aplicável
		Através de um algoritmo que conta o número de ocorrências dos termos da taxonomia no texto, é exibida uma lista em ordem decrescente para a adoção ou não dos descritores sugeridos.	A indexação é feita através de atribuição de termos semânticos definidos pelo indexador.
Geração de lista de elementos e suas referências no documento baseado no vocabulário controlado.	Gerar um grupo de termos que, baseados na Taxonomia, se encontram no texto.	Aplicado	Não aplicável
		A fim de escolher os melhores descritores com base na Taxonomia, uma listagem é apresentada ao indexador para a adoção ou não dos termos.	Os descritores (semânticos ou literais) não dependem de uma linguagem documentária.
Geração de metadados em padrão *Dublin Core*	Permitir a estruturação das informações em linguagens documentárias.	Aplicado	Não aplicável
		Para padronizar os metadados, pelo menos 1 (um) padrão foi implementado. Há ainda o objetivo futuro de adotar outros padrões.	
Permitir a pesquisa	Permitir a	Não implementado	Aplicado

no documento, mediante o mapa semântico, da linguagem documentária (taxonomia) e exibir o contexto no documento.	recuperação do item através de mapa semântico (gráfico) trazendo ao leitor, não apenas o item, mas o local dentro do documento ao qual o termo foi indexado.	Diante da dificuldade em criar uma interface gráfica em mapa semântico, optou-se por uma interface mais simples, textual, baseada na Taxonomia ou através de Linguagem Natural, além da listagem completa do acervo ordenada por critérios do usuário.	Através de *software* específico (*InxightStarTree*), a recuperação em contexto pode ser acessada pelo mapa semântico.

Fonte: Autor

Através desta reflexão, o trabalho afirma a importância dos metadados no ambiente das Bibliotecas Digitais, sendo seus principais objetivos:

a) possibilitar a estruturação de informações não estruturadas;
b) definir um padrão para entrada (cadastro) e saída (recuperação) das informações;
c) trabalhar a indexação (catalogação de assunto) e a recuperação através de um mesmo instrumento;
d) permitir a tradução da linguagem natural para a linguagem documentária;
e) implementar a interoperabilidade na medida em que outros *softwares* utilizarem o mesmo padrão através de padrões abertos de codificação.

Cabe também destacar alguns avanços significativos no processo de construção do protótipo, como:

a) visualização nativa do arquivo PDF dentro do ambiente do *software*;
b) extração dos metadados do próprio PDF, que, quando preenchidos, ajudam na indexação do item no acervo;
c) adoção do instrumento taxonômico, que pode ser criado, substituído ou mesmo editado pelo usuário, adicionando ou retirando termos;
d) possibilidade de trabalhar a indexação enquanto entrada e a seleção de termos/linguagem natural enquanto saída/recuperação, através de um mesmo ambiente para o profissional da informação;

e) manipulação dos metadados, taxonomia e configurações em padrão aberto XML, para permitir independência de *softwares* terceiros e possibilitar a compatibilização em outros sistemas;
f) recuperação através da linguagem natural, direcionado a pesquisa a metadados específicos, ou recuperação através da taxonomia utilizada na indexação, padronizando a recuperação, aumentando a assertividade e precisão no processo de busca;
g) auxilio ao profissional da informação e ao usuário final no processo de indexação através de um mecanismo semiautomático.

O protótipo alcançou em parte o objetivo projetado, e alguns desafios devem ser trabalhados a fim de concluir o modelo proposto pelo projeto MHTX. Podem ser elencados os seguintes processos para trabalhos futuros:

a) migração do modelo *Desktop* ou rede local para o modelo *Web*;
b) importação de outras Taxonomias gravadas em outros formatos;
c) atribuição de termos semânticos ao documento;
d) recuperação gráfica por mapas semânticos;
e) recuperação no contexto textual de partes específicas do documento e não apenas do item bibliográfico.

Sobre o problema inicial que norteou a pesquisa: *Como os metadados podem ser utilizados na gestão de uma biblioteca digital e impactar na recuperação de itens específicos no acervo?*, pode-se afirmar que os metadados foram de importância crucial para estabelecer a estratégia que possibilitou a fundamentação do protótipo. Os metadados apresentam-se, neste projeto, como a base na qual os demais processos foram desenvolvidos.

Conclui-se, pois, que os objetivos tiveram um alcance satisfatório frente à proposição do trabalho:

a) os metadados proporcionaram maior eficácia e assertividade, além de ajudar nas questões relacionadas à usabilidade do sistema;
b) o *software* desenvolvido é plenamente utilizável na versão do lançamento deste trabalho e possui funcionalidades de tratamento/entrada e recuperação/saída das informações de um acervo específico.

Este trabalho pretende proporcionar, no ambiente acadêmico, ensejo para que alunos e professores ampliem suas discussões sobre temas de grande relevância para a Ciência da Informação, como os vocabulários controlados, as linguagens documentárias, os processos de indexação (catalogação de assunto), os sistemas de recuperação da informação (SRI) e os metadados, além do viés tecnológico presente na manipulação do *software,* da representação informacional gravada através das linguagens XML/RDF e recuperada através de interfaces textuais e mapas hiperbólicos.

É certo também que este trabalho pode influenciar novas linhas de pesquisa relacionadas à área de organização e uso da informação, bem como as ligadas à área de tecnologia da informação. Novos projetos baseados neste protótipo poderão, por exemplo, trabalhar questões como:

a) a linguagem documentária (taxonomia) e seu papel na tradução e recuperação da informação em acervos digitais. Essa linha de pesquisa pode evidenciar, com mais detalhes, a importância dos instrumentos de organização e classificação no ambiente dos sistemas de recuperação da informação;
b) as possibilidades da indexação automática com base em instrumentos de organização e classificação da informação;
c) indexação semântica *versus* indexação literal: as características e consequências da tradução semântica, seus desafios através do polimorfismo e dos regionalismos;
d) interfaces ricas (*desktop*) *versus* interfaces *web*: as possibilidades de interação com o usuário;
e) o papel dos metadados e estruturas de organização em acervos controlados (específicos) e em ambientes não controlados (Internet, por exemplo), seus desafios e estratégias para a organização da informação.

Ao concluir este trabalho, afirmo que foi uma experiência realizadora na medida em que os fundamentos teóricos da academia puderam ser colocados em prática através da prototipação do *software* e suas aplicações práticas. Acredito que esta abordagem será uma contribuição válida para o ambiente acadêmico, apoiando diversas pesquisas atuais, inseridas no contexto deste objeto de estudo e seus desdobramentos, além de fomentar novas discussões e trabalhos.

O protótipo desenvolvido neste trabalho será divulgado para a utilização no ambiente educacional, por professores, alunos e escolas de qualquer natureza, privada ou estatal, em todos os níveis de ensino. A gratuidade desse software é uma forma de retornar

para a sociedade o investimento realizado no ensino público federal, que proporciona a tantas pessoas um ambiente de qualidade e produção intelectual. Isso se dará através de um produto tangível – o protótipo MeTa.

REFERÊNCIAS

ADIDA, Ben et al. *RDFA 1.1 Primer*: rich structured data markup for web documents. June 2012. (W3C Working Group Note 07). Disponível em: <http://www.w3.org/TR/ xhtml-rdfa-primer/>. Acesso em: 20 set. 2012.

ADOBE SYSTEMS. *Adobe XMP Developed Center*. Disponível em: <http://www. adobe.com/devnet/xmp.html>. Acesso em: 19 set. 2012.

ALEXANDER, Mary. S. *Core Cataloging and metadata standards and best practices*. Alabama:The Haworth Press, 2008.

ALVES, Maria das Dores Rosa; SOUZA, Márcia Izabel Fugisawa. Estudo de correspondência de elementos metadados: Dublin Core and Marc 21.

BECKETT, Dave. (Ed.) *RDF/XML syntax specification (revised)***.** 2004. Disponível em: <http://www.w3.org/TR/2004/REC-rdf-syntax-grammar-20040210/>. Acesso em: 20 set. 2012.

BIBLIOTECA NACIONAL (Brasil). *Biblioteca Nacional Digital Brasil*. Disponível em: <http://bndigital.bn.br/>. Acesso em: 19 set. 2012.

BORKO, Harold. *Indexing concepts and methods***.** New York: Academic Press, 1978. 261p.

BRASIL. Ministério da Educação. *Portal Domínio Público*. Disponível em: <http:// www.dominiopublico.gov.br/pesquisa/PesquisaObraForm.jsp>. Acesso em: 19 set. 2012a.

BRASIL. Senado Federal. *Biblioteca Digital do Senado Federal*. Disponível em: <http://www2.senado.gov.br/bdsf/>. Acesso em: 19 set. 2012b.

BRASIL. Senado Federal. *Royalty***.** Portal de notícias, 11 fev. 2012c. Disponível em: <http://www12.senado.gov.br/noticias/search?SearchableText=Royalty>. Acesso em: 19 set. 2012

BURKE, Peter. Problemas causados por Gutenberg: a explosão da informação nos primórdios da Europa Moderna. *Estudos Avançados*, São Paulo, v. 16, n. 44, p. 173-185, jan./abr. 2002. Disponível em: <http://www.scielo.br/pdf/ea/v16n44/v16n44 a10.pdf>. Acesso em: 17 set. 2012.

BUSH, Vannevar. As we may think. *Atlantic Monthly*, v. 176, n. 1, p. 101-108, 1945.

CAMPOS, Maria Luíza de Almeida. *A organização de unidades do conhecimento em hiperdocumentos*: o modelo conceitual como um espaço comunicacional para realização da autoria. 2001. 190f. Tese (Doutorado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, Escola de Comunicação e Artes, Rio de Janeiro.

CAMPOS, Maria Luíza de Almeida; GOMES, Hagar Espanha. Taxonomia e classificação: o princípio de categorização. *DamaGamaZero: Revista de Ciência da Informação*, v. 9, n. 4, ago. 2008. Disponível em: <http://www.dgz.org.br/ago08/ F_I_onum.htm>. Acesso em: 17 set. 2012.

COLEMAN, Anita S. *How to create, apply, and use metadata*: part II. Disponível em: <http://polaris.gseis.ucla.edu/gleazer/260_readings/Thurman.pdf>. Acesso em: 15 fev. 2011.

CONSELHO NACIONAL DE DESENVOLVIMENTO CIENTÍFICO E TECNOLÓGICO. *Plataforma Lattes*. Disponível em: <http://lattes.cnpq.br/>. Acesso em: 19 set. 2012.

CUNHA, Murilo Bastos da. Desafios na Construção de uma Biblioteca Digital. *CI. Inf. v. 1999*.

CUNHA, Murilo Bastos. Das bibliotecas convencionais às digitais: diferenças e convergências. *Perspectivas em Ciência da Informação*, Belo Horizonte, v. 13, n. 1, p. 2-17, jan./abr. 2008.

DACONTA, Michael C.; OBRST, Leo J.; SMITH, Kevin T.*The semantic web:* a guide to the future of sml, web services, and knowledge management. New York: Wiley, 2003. 312p.

DEITEL, H. M. *XML*: como programar. Porto Alegre: Bookman, 2003.

DIAS, Eduardo Wense. Contexto digital e tratamento da informação. *DataGamaZero: Revista de Ciência da Informação*, v. 2, n. 5, out. 2001. Disponível em: <http://www.dgz.org.br/out01/F_I_art.htm>. Acesso em: 17 set. 2012.

DUBLIN CORE METADATA INITIATIVE. *Dublin core qualifiers*. Disponível em: <http://dublincore.org/documents/2000/07/11/dcmes-qualifiers>. Acesso em 15 fev. 2011.

DUBLIN CORE METADATA INITIATIVE. *Dublin Core Metadata Element Set, Version 1.1: Reference Description***.** Disponível em: <http://dublincore.org/ documents/2000/07/11/dcmes-qualifiers>. Acesso em 22 set. 2012.

DYBDAHL, Lars. *GNU Gettext for Delphi, C++ Builder*. 11 nov. 2008. Disponível em: <http://dxgettext.po.dk/Home>. Acesso em: 20 set. 2012.

FEITOSA, Ailton. *Organização da informação na web***:** das tags à web semântica. Brasília: Thesaurus, 2006. (Estudos Avançados em Ciência da Informação, v. 2).

FOULONNEAU, Muriel; RILEY, Jenn. *Metadata for Digital Resources*: Implementation, systems disign and interoperability. Oxford: Chandos Publishing. 2008.

FRANÇA, Júnia Lessa et al. *Manual para normalização de publicações técnico-científica*s**.** 8. ed. Belo Horizonte: UFMG, 2007. 255p.

FUJITA (ORG.), Mariângela S. Lopes. *A Indexação de livros*: A percepção de catalogadores e usuários de bibliotecas universitárias. São Paulo: Ed. Cultura Acadêmica – UNESP, 2009.

GONÇALVES, Maria Carolina. A Percepção de Usuários sobre a Indexação na Análise de Assuntos para Catalogação. Dissertação de Mestrado. 2009. In: FUJITA (ORG.), Mariângela S. Lopes. *A Indexação de livros*: A percepção de catalogadores e usuários de bibliotecas universitárias. São Paulo: Ed. Cultura Acadêmica – UNESP, 2009.

GRANITZER, Michael; LUX, Mathias; SPANIOL, Marc. *Multimedia Semantics*: The Role of Metadata. Studies in Computacional Inteligence. Berlin: Springer. 2008

HAWKINS, Donald T; LARSON, Signe E.; CATON, Bari Q. Information science abstracts: tracking the literature of information science. part 2: a new taxonomy for information science. *Journal of the American Society for Information Science and Technology***,** v. 54, n. 8, p.77-781, 2003.

INGO SCHMOEKEL. *Pdf-analyser.com*. Disponível em: <http://www.pdf-analyzer.com>. Acesso em: 24 set. 2012.

INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA E TECNOLOGIA. *Workshop para implantação de biblioteca digital de teses e dissertações na Universidade Metodista de São Paulo e na Universidade Presbiteriana Mackenzie*. São Paulo: IBICT, 2006. Disponível em: <http://tedesite.ibict.br/BDTD_ Workshop_Abril2006.PPT#375,1>. Acesso em: 19 set. 2012.

INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA E TECNOLOGIA. *Biblioteca Digital Brasileira de Teses e Dissertações*. Disponível em: <http://bdtd.ibict.br/>. Acesso em: 19 set. 2012.

LAFON, Yves; BOS, Bert. *Describing and retrieving photos using RDF and HTTP*. W3C Notes, 19 April 2002. Disponível em: <http://www.w3.org/TR/photo-rdf/>. Acesso em: 17 set. 2012.

LANCASTER, F. W. *Indexação e resumos***:** teoria e prática. 2. ed. Brasília: Briquet de Lemos Livros, 2004. 452p.

LEVY, Pierre. Os três tempos do espírito: a oralidade primária, a escrita e a informática. In: LEVY, Pierre. *As tecnologias da inteligência*: o futuro do pensamento na era da informática. Rio de Janeiro: Ed. 34, 1993. Cap. 2, p. 75-132.

LIMA, G. A. B. O. Mapa conceitual como ferramenta para organização do conhecimento em sistemas de hipertextos e seus aspectos cognitivos. *Perspectivas em Ciência da Informação.* Belo Horizonte: v. 9, n. 2, p. 134-145, 2004a.

LIMA, G. A. B. O. (Coord.) *Protótipo Mapa Hipertextual - MHTX*: um modelo para organização hipertextual de documentos acadêmicos por meio do uso de mapas conceituais, analise facetada e sistemas hipertextuais. Belo Horizonte: Escola de Ciência da Informação da UFMG. Disponível em: <http://www.gercinalima.com/ mhtx/>. Acesso em: 24 set. 2012**.**

LIMA, G. A. B. O. *Mapa hipertextual (MHTX):* um modelo para organização hipertextual de documentos. 2004b. 199f. Tese (Doutorado em Ciência da Informação) - Universidade Federal de Minas Gerais, Escola de Ciência da Informação, Belo Horizonte.

LIMA, Vania Mara Alves; BOCCATO, Vera Regina Casari. O Desempenho Terminologico dos Descritores em Ciência da Informação do Vocabulário Controlado do SIBi/USP nos Processos de Indexação Manual, Automática e Semi-automática. *Perspectiva em Ciência da Informação.* São Paulo*:* v. 14, n. 1, p. 131-151. Jan/abr. 2009.

LIU, Jia. *Metadata and its applications in the digital library:* approaches and practices. Westport: Libraries Unlimited, 2007. 212p.

LOURENÇO, Cíntia. Azevedo. *Análise do padrão brasileiro de metadados de teses e dissertações segundo o modelo entidade-relacionamento*. 2004. Tese (Doutorado em Ciência da Informação) - Universidade Federal de Minas Gerais, Escola de Ciência da Informação, Belo Horizonte.

LOURENÇO, Cíntia Azevedo. Metadados: o grande desafio na organização da web. *Informação & Sociedade: Estudos*, João Pessoa, v. 17, n. 1, p. 71-80, jan./abr. 2007. Disponível em: <http://www.ies.ufpb.br/ojs2/index.php/ies/article/view/466/ 1466>. Acesso em: 17 set. 2012.

MARCONDES, Carlos H. et al. (Org.). *Bibliotecas digitais:* saberes e práticas. Brasília: IBICT, 2006. 278p.

MARCONI, Maria de Andrade; LAKATOS, E. M. *Fundamentos da metodologia científica*. 5. ed. São Paulo: Atlas, 2003.

MAURER, Margareth Beecher, NICKERSON, Joshua. Morphing metadata: maximizing access to eletronic theses and dissertations. *Liberty and Tech*. v. 20, n. 1, p. 11-57 Emerald. 2008

MICROSOFT. *XML Notepad 2007*. Disponível em: <http://www.microsoft.com/ download/en/details.aspx?id=7973>. Acesso em: 19 set. 2012.

MORAES, Alice Ferry de; OLIVEIRA, Telma Marisa de. Experiências Relacionadas ao Levantamento de Teses e Dissertações. *Inf. de Soc: Est*. João Pessoa: v. 20, n 1, p. 73- 81., jan/abr. 2010.

NAVES, Madalena Martins Lopes; KURAMOTO, Hélio. *Organização da informação***:** princípios e tendências. Brasília: Briquet de Lemos, 2006.

NAVES, Madalena Martins Lopes. Estudo de fatores interferentes no processo de análise de assunto. *Perspect. Cienc. Inf*. Belo Horizonte: v. 6, n. 2, p. 189-203, jul/dez. 2001.

NETWORKED DIGITAL LIBRARY OF THESES AND DISSERTATIONS. *About NDLTD*. Disponível em: <http://www.ndltd.org/about>. Acesso em: 19 set. 2012.

ODDONE, N. E.; GOMES, M. Y. F. S. Os temas de pesquisa em ciência da informação e suas implicações político-epistemológicas. In: ENCONTRO NACIONAL DE CIÊNCIA DA INFORMAÇÃO, 5, 2004, Salvador. *Anais...* Brasília: IBICT, 2004. Disponível em: <http://dici.ibict.br/archive/00000558/01/temas_de_pesquisa.pdf> Acesso em: 08 ago, 2009.

ONLINE COMPUTER LIBRARY CENTER. *Dublin Core releases recommended qualifiers to improve access to information***.** Jul. 2000. Disponível em: <http://www.oclc.org/research/news/2000/07-21b.html>. Acesso em: 19 set. 2012.

OPEN ARCHIVES. *The Open Archives initiative protocol for metadata harvesting*: protocol version 2.0. 2002. Disponível em: <http://www.openarchives.org/ OAI/openarchivesprotocol.html>. Acesso em: 19 set. 2012.

ROBREDO, Jaime. *Documentação de hoje e de amanhã***:** uma abordagem revisitada e contemporânea da Ciência da Informação e de suas aplicações biblioteconômicas, documentárias, arquivísticas e museológicas. 4. ed. Brasília: Reproart, 2005. 409 p.

ROCHA, Rafael Port da. Metadados, web semântica, categorização automática: combinando esforços humanos e computacionais para a descoberta e uso dos recursos da web. *Em Questão***:** Revista da Faculdade de Biblioteconomia e Comunicação da UFRGS**,** Porto Alegre, v. 10, n. 1, p. 109-121, jan./jun. 2004. Disponível em: <http://www.seer.ufrgs.br/index.php/EmQuestao/article/viewArticle/ 86>. Acesso em: 9 abr. 2009.

ROSETTO, Márcia. Bibliotecas digitais: cenário e perspectivas. *Revista Brasileira de Biblioteconomia e Documentação*, São Paulo, v. 4, n. 1, p. 101-130, jan./jun. 2008.

ROSETTO, Márcia. *Metadados e recuperação da informação*: padrões para bibliotecas digitais. Trabalho apresentado na II Cibernética: Simpósio Internacional de Propriedade Intelectual, Informação e Ética; VIII Encontro Nacional de Informação e Documentação Jutídica; XXII Painel Biblioteconomia em Santa Catarina. Florianópolis, SC, 12 a 14 de novembro de 2003. Disponível em: <http://www.sibi. usp.br/sibi/boletim_inter/vol_8_num_6/SIBICiberetica.doc>. Acesso em: 17 set. 2012.

RUBI, Milena Polsinelli. Os Princípios da Política de Indexação na Análise de Assunto para catalogação: especificidade, exaustividade, revocação e precisão na perspectiva dos catalogadores e usuários. In: FUJITA (org.), Mariângela Spotti Lopes. *A Indexação de Livros*: a percepção de catalogadores e usuários de bibliotecas universitárias. Ed. Cultura Acadêmica UNESP. São Paulo: 2009.

SARACEVIC, Tefko. Ciência da informação: origem, evolução e relações. *Perspectivas em Ciência da Informação.* Belo Horizonte: v. 1, n. 1, p. 41-62, jan./jun. 1996.

SAYAO, Luis Fernando; MARCONDES, Carlos Henrique. O desafio da interoperabilidade e as novas perspectivas para as bibliotecas digitais. *TransInformação***,** Campinas, v. 20, n. 2, p. 133-148, maio/ago. 2008.

SICILIA, Miguel-Angel; LYTRAS, Miltiadis D. *Metadata and Semantics*. Springer. New Yourk: 2009.

SOERGEL, Dagobert. Digital Libraries and Knowledge Organization. Sebastian Ryszard Kruk and Bill McDaniel. Eds. *Semantic Digital Libraries*. 2008.

SOUTHWICK, Silvia Barcellos. *Biblioteca Digital Brasileira de Teses e Dissertações*: modelo e tecnologias. Brasília: IBICT, 2003. Disponível em: <http://bdtd.ibict.br/images/stories/documentos_importantes/bdtd_documentosilvia.doc>. Acesso em: 17 set. 2012.

SOUZA, Renato Rocha; ALVARENGA, Lídia. A web semântica e suas contribuições para a ciência da informação. *Ciência da Informação*, Brasília, v. 33, n. 1, p. 132-141, jan./abr. 2004.

UNIVERSIDADE DE CAMPINAS. *Biblioteca Digital da UNICAMP*. Disponível em: <http://cutter.unicamp.br/>. Acesso em: 19 set. 2012.

UNIVERSIDADE DE SÃO PAULO. *Biblioteca Digital de Teses e Dissertações da USP*. Disponível em: <http://www.teses.usp.br/>. Acesso em: 19 set. 2012.

UNIVERSIDADE FEDERAL DE MINAS GERAIS. *Biblioteca Digital*. Disponível em: <http://www.lcc.ufmg.br/index.php?option=com_content&view=article&id=14&Itemid=15>. Acesso em: 17 set. 2012.

VIRGINIA TECH. *Networked Digital Library of Theses Dissertations*. Disponível em: <http://scholar.lib.vt.edu/theses/ndltd.html>. Acesso em: 19 set. 2012.

WORLD WIDE WEB CONSORTIUM. *What is linked data?* Disponível em: <http://www.w3.org/standards/semanticweb/data>. Acesso em: 19 set. 2012b.

APÊNDICES

Apêndice A – Elementos do Dubin Core

Como ler o quadro abaixo: http://dublincore.org/documents/1999/07/02/dces/

Elemento	title (título)
Rótulo	Title
Definição	O nome dado a um determinado recurso.
Comentário	Tipicamente, o título será o nome que o recurso será formalmente reconhecido.
Valor controlado	Não
Notas	Geralmente, o título consegue expressar o significado do recurso. Todavia pode não haver uma distinção clara em alguns recursos eletrônicos quando sua descrição é genérica.

Elemento	creator (criador)
Rótulo	Creator
Definição	A entidade primária responsável (em primeira instância) pela criação do conteúdo do recurso.
Comentário	Como exemplos do Criador podemos citar: uma pessoa, uma organização ou um serviço. Tipicamente o nome de um Criador deve ser usado para indicar uma entidade.
Valor controlado	Não
Notas	Catalogar o Criador pode ser uma atividade subjetiva, pois a referência pode não estar listada junto ao recurso. Embora seja possível despender tempo extra para procurar esta informação geralmente isso leva muito tempo. Se o criador não pode ser identificado facilmente, deixe este campo em branco.

Elemento	subject (assunto)
Rótulo	Subject
Definição	Tópico do conteúdo do recurso.
Comentário	Tipicamente, o Assunto deve ser expresso como palavras chave, frases chave ou códigos de classificação que descrevem um tópico do recurso. Recomenda-se como melhores práticas, a adoção de termos retirados de vocabulários controlados ou esquemas de classificação.
Valor controlado	Opcional. Pode-se utilizar: LCSH, MeSH, DDC, LCC e UDC. Ou palavras chave do próprio recurso.
Notas	É objetivo quando os termos são encontrados no recurso. É subjetivo quando o catalogador sugere um termo não encontrado no recurso. Na maioria das vezes o catalogador fornece essa informação

Elemento	description (descrição)
Rótulo	Description
Definição	Uma descrição do conteúdo do recurso.
Comentário	Descrição pode incluir (mas não está limitado a): um resumo, tabela de conteúdo, referência a uma representação gráfica do conteúdo ou um texto livre sobre o conteúdo.

Valor controlado	Não
Notas	A informação é objetiva se o conteúdo for extraído do recurso. Será subjetiva se for fornecida pelo catalogador.

Elemento	publisher (editor)
Rótulo	Publisher
Definição	Entidade responsável por tornar o recurso acessível.
Comentário	Exemplos para Editor incluem uma pessoa, uma organização ou um serviço. Tipicamente o nome de um Editor deve ser usado para indicar uma entidade.
Valor controlado	Não
Notas	É difícil encontrar o Editor em publicações de recursos eletrônicos. Por exemplo, na publicação de uma página na web, quem é o Editor? Geralmente consideramos a organização ou indivíduo que publicou o arquivo no servidor web como sendo o Editor, neste raciocínio o Editor para a página web seria a Universidade do Arizona.

Elemento	contributor (colaborador)
Rótulo	Contributor
Definição	Responsável por registrar os colaboradores do conteúdo do recurso.
Comentário	Exemplos de colaboradores incluem: Uma pessoa, uma organização ou um serviço.
Valor controlado	Não
Notas	Há muitas maneiras de definir um colaborador, mas a fim de ganhar tempo e optar pela simplicidade a escolha de três níveis de colaboração são suficientes.

Elemento	date (data)
Rótulo	Date
Definição	A data associada com o evento de ciclo de vida do recurso.
Comentário	Tipicamente Date será associado com a criação ou disponibilidade do recurso.
Valor controlado	Opcional. É uma recomendação de boa prática a utilização dos padrões de codificação descritos na norma ISO 8601 [W3CDTF] e a configuração de formato YYYY-MM-DD.
Notas	Se não há como definir Date, deixe em branco. É possível usar os seguintes qualificadores: - *Created* (Criação do recurso) - *Valid* (Validação do recurso) - *Available* (Disponibilização do recurso) - *Issued* (Publicação do recurso) - *Modified* (Alteração do recurso) - *Accepted* (Aceitação do recurso) - *Submitted* (Apresentação do recurso)

Elemento	type (tipo)
Rótulo	Type
Definição	A natureza ou gênero do conteúdo do recurso.
Comentário	O tipo inclui termos de que descrevem categorias de forma geral, funções, gêneros ou níveis de agregação para o conteúdo. É recomendado como melhor prática a seleção destas terminologias extraídas de um vocabulário controlado.
Valor controlado	Sim. Existe uma lista de 10 tipos usados através dos tipos do vocabulário DCMI:

	- *Collection*: Uma coleção é uma agregação de itens. Quer dizer que o recurso é descrito como um grupo. - *Dataset*: É uma informação codificada em formato estruturado bem definido como exemplo: listas, tabelas e banco de dados. - *Event*: Um evento é uma ocorrência não persistente, baseada em um espaço de tempo. Exemplos: conferências, workshops, cerimônias. - *Image*: A imagem é considerada em primeiro lugar, um símbolo visual e uma manifestação informacional além do texto. - *Interactive Resource*: Um recurso interativo é um tipo de recurso que requer interação do usuário para ser compreendido, executado ou vivenciado. Por exemplo: Formulários na Internet, objetos multimídia para aprendizado, serviços de comunicação e realidade virtual. - *Service*: Um serviço é um sistema que provê uma ou mais funções de valor para o usuário final. Exemplos: serviços de fotocópia, Internet banking, autenticação eletrônica. - *Software*: É um programa de computador na forma de código fonte ou na forma executável (utilização final). - *Sound*: Som é um recurso que contém primariamente a intenção de comunicação através do sentido auditivo. - *Text*: É um recurso que contém primariamente informação textual. - *Physical Object*: É um objeto inanimado, tridimensional ou substancial. Como exemplos: um computador, uma pirâmide ou uma escultura. Devemos notar que a representação digital ou substitutos para estes elementos, podem usar os tipos *Image*, *Text* ou outros.
Notas	

Elemento	format (formato)
Rótulo	Format
Definição	É a manifestação física ou digital do recurso.
Comentário	Tipicamente o formato inclui o tipo de mídia ou as dimensões do recurso. Pode ser usado para determinar se é necessário um software, hardware ou outro equipamento para apresentar ou operar o recurso. Exemplos de dimensão incluem tamanho e duração. É recomendado como melhor prática a utilização de valores extraídos de vocabulário controlado, exemplo: Internet Assigned Numbers Authority - IANA onde os tipos de arquivos são especificados através da lista de tipos MIME.
Valor controlado	Sim. É uma boa prática optar pela utilização de termos registrados na IANA.
Notas	Alguns tipos mais populares são: text/html text/xml text/rtf

Elemento	identifier (identificador de recurso)
Rótulo	Identifier
Definição	Uma referência não ambígua (única) a fim de especificar o recurso em um determinado contexto.
Comentário	É recomendado identificar o recurso através de códigos textuais ou numéricos conforme o nível informacional do sistema. Exemplos: na Internet a URL/URI e a ISBN no contexto publicações.
Valor controlado	Não. Mas há a necessidade de respeitar o padrão escolhido (URL ou ISBN, etc...)

Notas	

Elemento	source (fonte)
Rótulo	Source
Definição	Indica a referência da qual o recurso foi derivado.
Comentário	O recurso atual pode ter sido derivado de uma fonte de modo inteiro ou parcial.
Valor controlado	Não.
Notas	Pode ser usado identificando o título seguido pela URL ou ainda descrito com suas próprias palavras (linguagem natural).

Elemento	language (língua)
Rótulo	Language
Definição	Linguagem intelectual do conteúdo do recurso.
Comentário	É recomendado como melhor prática a adoção de termos extraídos da norma RFC 3066 (http://www.ietf.org) que em conjunto com a ISO 632.2 (http://www.loc.gov/standards/iso639-2/php/code_list.php) definem códigos em dois ou três dígitos para determinar a linguagem, exemplo: “en” ou “eng” para a língua Inglesa.
Valor controlado	Não. Mas recomenda-se a adoção de um padrão.
Notas	

Elemento	relation (relação)
Rótulo	Relation
Definição	Permite documentar a referência a um recurso relacionado.
Comentário	É recomendado como boa prática para representar a relação com outro recurso, o código textual ou numérico conforme o nível informacional do sistema.
Valor controlado	Não. É possível implementar os seguintes refinamentos: - *IsVersionOf*: - *HasVersion*: - *IsReplacedBy*: - *Replaces*: - *IsRequiredBy*: - *Requires*: - *IsPartOf*: - *HasPart*: - *IsReferencedBy*: - *References*: - *IsFormatOf*: - *HasFormat*: - *ConformsTo*:
Notas	

Elemento	coverage (cobertura)
Rótulo	Coverage
Definição	Descreve a extensão do escopo do conteúdo do recurso.
Comentário	Este identificador será usado tipicamente para incluir informações geográficas e históricas em termos de localização espacial. Exemplos: coordenadas geográficas, lugares e endereços.
Valor controlado	Não
Notas	

Elemento	rights (direitos)

Rótulo	Rights
Definição	É a informação dos direitos sobre o recurso.
Comentário	Tipicamente é usado para registrar os direitos autorais ou estatutos que podem determinar o provedor/responsável/proprietário do recurso. Marcas registradas e Direitos de propriedade intelectual podem ser registrados neste elemento através de uma lista de termos.
Valor controlado	Sim. A lista é composta por: - Accessible freely - Licence restrictions apply - Restrictions apply - Subscription needed - Public domain
Notas	

Elemento	audience (audiência)
Rótulo	Audience
Definição	Permite qualificar a pretensão do público de usuários para o recurso.
Comentário	Se não preenchido, o próprio usuário deve julgar a pertinência ou não do recurso em seus interesses.
Valor controlado	Sim. A lista é composta por: - Elementary: - Middle School: - High School: - Undergraduate Level: - Graduate Level: - Professional: - General Education:
Notas	É um elemento subjetivo.

ANEXOS

Anexo A – Taxonomia da Ciência da Informação

Nanci Oddone e Maria Yêda F. S. de Filgueiras Gomes

01 – Aspectos teóricos e gerais da ciência da informação

Bibliometria, cienciometria, infometria

Biblioteconomia comparada

Biblioterapia

Conceitos de biblioteca

Ética e ciência da informação

Fundamentação epistemológica

História da arquivologia, da biblioteconomia, da documentação e da ciência da informação

História do livro e das bibliotecas

Interdisciplinaridade

Leis bibliométricas

Metodologia da pesquisa

Origem e evolução da ciência da informação

Pesquisa científica

Teoria dos sistemas

Teorias e conceitos de informação

Outras questões teóricas

02 – Formação profissional e mercado de trabalho

Avaliação de cursos

Currículo, metodologia e programa de ensino

Formação profissional

Profissional da informação

Profissões e mercado de trabalho

03 – Gerência de serviços e unidades de informação

Arquivos
Automação de unidades de informação
Avaliação de bases de dados
Avaliação e desenvolvimento de coleções
Avaliação de serviços e de unidades de informação
Balcão de informações
Consórcios
Compartilhamento de recursos
Comportamento gerencial
Custos
Estilos gerenciais
Gerência de recursos informacionais (GRI)
Gerência organizacional
Gestão da qualidade
Inteligência competitiva
Marketing
Monitoramento ambiental
Motivação
Pesquisa de mercado
Planejamento, organização e gerência de serviços e de unidades de informação
Processo decisório
Recursos financeiros
Recursos humanos
Serviços de extensão bibliotecária
Sistemas de informação gerencial
Estudos sobre outros serviços e unidades de informação

04 – Estudos de usuário, demanda e uso da informação e de unidades de informação

Caracterização e comportamento do usuário
Educação e treinamento de usuários
Hábitos de leitura
Necessidades de informação
Oferta, demanda e transferência de informação
Uso e impacto das novas tecnologias de comunicação e informação
Usos da informação e de unidades de informação

05 – Comunicação, divulgação e produção editorial

Atividade editorial
Avaliação de periódicos
Divulgação científica
Documentação científica
Editoração/publicação eletrônica
Estudos bibliométricos, cienciométricos e infométricos
Estudos da produção e da produtividade científica
Estudos de autoria
Estudos de canais, veículos, ciclos e modelos de comunicação
Estudos de citação
Estudos sobre fontes de informação
Indicadores de produtividade científica
Jornalismo científico
Literatura cinzenta
Normalização
Produção editorial de impressos
Produção do texto científico
Publicações oficiais

06 – Informação, cultura e sociedade

Alfabetização digital
Biblioteca, cultura e sociedade
Centros populares de documentação e comunicação
Democratização da informação
Inclusão/exclusão informacional
Informação, ação cultural e cidadania
Sociedade da informação

07 – Legislação, políticas públicas de informação e de cultura

Depósito legal
Direitos de propriedade intelectual
Economia da informação
Indústria e mercado cultural
Indústria e mercado da informação
Informação ambiental
Informação científica e tecnológica
Informação para indústria e negócios
Informação tecnológica
Política científica e tecnológica
Política cultural
Política de informação
Política de informação científica e tecnológica
Política editorial
Transferência de tecnologia

08 – Tecnologias da informação

Bases de dados
Bibliotecas virtual, digital e eletrônica
CD-ROM
Hipertexto e hipermídia
Mecanismos de busca (*search engines*)
Redes eletrônicas de informação
Sistemas de gerenciamento eletrônico de documentos (GED)
Sistemas especialistas
Sistemas para automação de unidades de informação
Tecnologias de inteligência competitiva
Outros sistemas e tecnologias de comunicação e informação

09 – Processamento, recuperação e disseminação da informação

Análise documentária
Catalogação/catalogação cooperativa
Classificação
Controle bibliográfico
Desenvolvimento de coleções
Elaboração de resumos

Indexação (manual e automática)
Linguagens documentárias
Normalização
Metadados
Preservação e conservação
Retirada e descarte
Recuperação da informação
Seleção e aquisição
Tesauros
Videotexto

10 – Assuntos correlatos e outros

Análise do discurso
Arquitetura da informação
Comunicação social
Design da informação
Informática
Lingüística
Telecomunicações