UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
FACULDADE DE BIBLIOTECONOMIA E COMUNICAÇÃO
DEPARTAMENTO DE CIÊNCIAS DA INFORMAÇÃO

ARTHUR POCHMANN MULINARI

MODELOS DE INTEROPERABILIDADE EM FERRAMENTAS DE CATALOGAÇÃO PARA COLEÇÕES PESSOAIS DE MÍDIA DIGITAL

Porto Alegre
2012

ARTHUR POCHMANN MULINARI

MODELOS DE INTEROPERABILIDADE EM FERRAMENTAS DE CATALOGAÇÃO PARA COLEÇÕES PESSOAIS DE MÍDIA DIGITAL

Trabalho de Conclusão de Curso apresentado como requisito parcial para obtenção do título de Bacharel em Biblioteconomia da Faculdade de Biblioteconomia e Comunicação da Universidade Federal do Rio Grande do Sul.

Orientador: Prof. Dr. Rafael Port da Rocha

Porto Alegre
2012

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
Reitor: Prof. Dr. Carlos Alexandre Netto
Vice Reitor: Prof. Dr. Rui Vicente Oppermann

FACULDADE DE BIBLIOTECONOMIA E COMUNICAÇÃO
Diretor: Prof. Ricardo Schneiders da Silva
Vice-diretora: Profa. Dra. Regina Van der Lann

DEPARTAMENTO DE CIÊNCIAS DA INFORMAÇÃO
Chefe: Profa. Dra. Ana Maria Mielniczuk de Moura
Chefe-substituta: Profa. Dra. Sônia Elisa Caregnato

COMISSÃO DE GRADUAÇÃO DA BIBLIOTECONOMIA
Coordenadora: Profa. Dra. Samile Andréa de Souza Vanz
Coordenadora Substituta: Profa. Ma. Glória Isabel Sattamini Ferreira

CIP. Brasil. Dados Internacionais de Catalogação na Publicação

S957p Mulinari, Arthur Pochmann
Modelos de Interoperabilidade em Ferramentas de Catalogação para Coleções Pessoais de Mídia Digital / Arthur Pochmann Mulinari – 2012.
85 f. : il. ; 30 cm.

Trabalho de Conclusão de Curso (Graduação) - Universidade Federal do Rio Grande do Sul. Faculdade de Biblioteconomia e Documentação. Curso de Biblioteconomia, 2012.

Orientador: Prof. Dr. Rafael Port da Rocha

1. Interoperabilidade 2. Mídia Digital 3. Coleções Pessoais I. Rocha, Rafael Port da II. Título.

CDU 027.1

Departamento de Ciências da Informação
Rua Ramiro Barcellos, 2705, sala 507
CEP: 90.035-007 – Porto Alegre/RS
Tel: (51) 3316.5143
Fax: (51) 3316.5435
E-mail: dci@ufrgs.br

Arthur Pochmann Mulinari

MODELOS DE INTEROPERABILIDADE EM FERRAMENTAS DE CATALOGAÇÃO PARA COLEÇÕES PESSOAIS DE MÍDIA DIGITAL

Monografia apresentada como requisito parcial para a obtenção do título de Bacharel em Biblioteconomia, da Faculdade de Biblioteconomia e Comunicação da Universidade Federal do Rio Grande do Sul.

Examinado em ..... de ..... de .....

Banca Examinadora

_____

Prof. Dr. Rafael Port da Rocha - Orientador
Departamento de Ciências da Informação - UFRGS

_____

Profa. Dra. Maria do Rocio Fontoura Teixeira
Departamento de Ciências da Informação - UFRGS

_____

Me. Gustavo Menezes Oliveira
Unidade de Infraestrutura e Tecnologia - Banrisul

Ao amor da minha vida, Marioni
– que não me deixou desistir.

RESUMO

Este trabalho tem o objetivo de analisar modelos de interoperabilidade adotados por ferramentas que automatizam a catalogação de coleções pessoais de músicas, vídeos e livros digitais. A metodologia empregada é uma combinação da pesquisa bibliográfica com a documental. Utiliza como instrumento de coleta dos dados o formulário, o qual aplica sobre três *softwares*: Clementine, XBMC e Calibre. Constata pela revisão de literatura e de documentos que os componentes do processo automatizado são três: a coleção digital, a fonte de informações *on-line* e os modelos de interoperabilidade. Classifica tais modelos em: busca federada, colheita e raspagem. Conclui, pela análise dos dados, que, no âmbito não tradicional das coleções pessoais, os modelos de interoperabilidade são implementados através de soluções de baixo custo não padronizadas, transferindo dados através da interface homem-máquina – característica da raspagem – sem garantias de disponibilidade da interoperabilidade com as fontes de informação *on-line*. Recomenda, aos desenvolvedores desse tipo de software, implementar os elementos faltantes do ISBD, salvar sempre os novos metadados descritivos diretamente nos arquivos, integrar a ferramenta às bases de dados de referência on-line.

PALAVRAS-CHAVE: Interoperabilidade. Mídia Digital. Coleções Pessoais. Metadados. Bases de dados.

ABSTRACT

This work aims to analyze interoperability models adopted by tools that automate the personal collection of video, music and digital book's cataloguing. The methodology used is a combination of literature with the documentary research. It uses a form as an instrument of data collection, which it applies on three softwares: Clementine, XBMC and Calibre. It notes by the literature and document review that the components of the automated process are three: a digital collection, the source of online information and the interoperability model. It classifies these models: federated search, harvesting and scraping. It concludes, by the data analysis, that under non-traditional personal collections the interoperability models are implemented through low-cost non standard solutions, transferring data through the man-machine interface – characteristic of scraping – without assuring availability of interoperability with the sources of online information. It recommends to the developers of such softwares to implement the missing elements of the ISBD, always save the new files directly in the descriptive metadata and to integrate the tool with the online reference databases.

KEYWORDS: Interoperability. Digital Media. Personal Collections. Metadata. Databases.

LISTA DE FIGURAS

Figura 1 – Níveis de abstração de uma base de dados 19
Figura 2 – Pasta de músicas classificadas em vários gêneros musicais 38
Figura 3 – Subpasta do gênero Rock com músicas classificadas por artistas.....39
Figura 4 – Biblioteca de música digital.....40
Figura 5 – Tela inicial de um *media center* 46
Figura 6 – Descrição de música no Clementine 63
Figura 7 – Descrição de livro no Calibre 64
Figura 8 – Descrição de filme no XBMC 65

LISTA DE QUADROS

Quadro 1 – Relação dos itens do formulário com os objetivos 56
Quadro 2 – Identificação dos *softwares*.....57
Quadro 3 – Fontes de informação para catalogação.....59
Quadro 4 – Modelo de interoperabilidade por ferramenta.....61
Quadro 5 – Elementos do ISBD que faltam em cada ferramenta 66
Quadro 6 – Método de identificação dos itens da coleção.....66
Quadro 7 – Local de salvamento dos novos metadados.....67

LISTA DE SIGLAS

ANSI …….. American National Standards Institute
API …… Application Programming Interface
CDDB……. Compact Disc Data Base
HTTP …….. Hypertext Transfer Protocol
IMDB ……... Internet Movie Database
ISAN ……..….. International Standard Audiovisual Number
ISBD …….. International Standard for Bibliographic Description
ISBN …….. International Standard Book Number
ISRC …….. International Standard Recording Code
NISO ……..…….. National Information Standards Organization
OAI ……..…… Open Archives Initiative
PMH ……..…… Protocol for Metadata Harvesting
SMPTE ……..….. Society of Motion Picture and Television Engineers
SPARCS …… Standards Planning and Requirements Committee
TheTVDB ……... The Television Database

SUMÁRIO

1 INTRODUÇÃO 11

1.1 IDENTIFICAÇÃO E JUSTIFICATIVA DO PROBLEMA 12

1.2 OBJETIVOS 14

1.2.1 Objetivo geral 14

1.2.2 Objetivos específicos 14

2 REVISÃO DE LITERATURA 16

2.1 BASES DE DADOS 17

2.1.1 Níveis de abstração 18

2.1.2 Categorização 20

2.1.3 Metadados 22

2.2 MODELOS DE INTEROPERABILIDADE 23

2.2.1 Busca federada de informações 26

*2.2.2 Colheita (*harvesting*) de metadados 27*

*2.2.3 Raspagem (*scraping*) de dados 29*

2.3 PRODUÇÃO DE METADADOS 31

2.3.1 Catalogação tradicional 32

2.3.2 Catalogação social 33

2.4 COLEÇÕES PESSOAIS DE MÍDIA DIGITAL 36

3 REVISÃO DE DOCUMENTOS 42

3.1 FERRAMENTAS DE CATALOGAÇÃO DE MÍDIA DIGITAL 42

3.1.1 Ferramenta para catálogos de músicas: Clementine 44

3.1.2 Ferramenta para catálogos de vídeos: XBMC 46

3.1.3 Ferramenta para catálogos de livros: Calibre 48

3.2 BASES UTILIZADAS PELAS FERRAMENTAS DE CATALOGAÇÃO 49

3.2.1 Bases sobre músicas: Gracenote e MusicBrainz 50

3.2.2 Bases sobre filmes e seriados: IMDB e TheTVDB 51

3.2.3 Bases sobre livros: LibraryThing e Anobii 51

4 METODOLOGIA 53

4.1 NATUREZA E ABORDAGEM 53

4.2 MÉTODO DA PESQUISA 53

4.3 OBJETO DA PESQUISA 55
4.4 INSTRUMENTO DE COLETA DOS DADOS 55

5 APRESENTAÇÃO E ANÁLISE DAS FERRAMENTAS 57

5.1 IDENTIFICAÇÃO DOS *SOFTWARES* 57
5.2 MEIOS DE INTEROPERABILIDADE 58
5.2.1 Fontes de informação utilizadas para catalogação 58
5.2.2 Nível de funcionalidades e custos do modelo de interoperabilidade 60
5.2.3 Adequação do formato dos metadados ao ISBD.....62
5.2.4 Método de identificação dos itens da coleção.....66
5.2.5 Local onde são salvos os novos metadados.....67

6 CONSIDERAÇÕES FINAIS.....68

REFERÊNCIAS.....71

1 INTRODUÇÃO

Não há precedentes na história humana de um período em que tanta informação estivesse disponível a tantas pessoas. Graças à popularização dos computadores e da rede mundial que os conecta, a Internet, praticamente qualquer pessoa pode agregar em uma coleção pessoal mais obras do que uma biblioteca pública poderia oferecer em suas estantes. As gerações mais novas, que já nasceram na era digital, podem não conseguir entender a real magnitude desse fato. Por isso, talvez, seja importante recordar, a título de exemplo, a trajetória de duas antigas coleções pessoais.

Conforme narra Bryson (2010), em 1757, Thomas Jefferson herdou quarenta e dois livros de seu pai. Esta quantia pode não parecer muito atualmente, mas, naquela época, livros não eram objetos tão comuns. Para efeito de comparação, quando John Harvard faleceu, doando a quantia de quatrocentos livros para uma recém criada universidade, ela foi rebatizada em sua homenagem. Thomas Jefferson (1743-1826) e John Harvard (1607-1638) viveram em séculos diferentes, por isso também era diferente o ritmo de crescimento de suas coleções. Harvard adquiria cerca de doze livros por ano, enquanto que Jefferson comprava, em média, a mesma quantia, mas a cada mês. Nesse ritmo, em 1815, este último possuía seis mil quatrocentos e quinze livros, os quais foram repassados ao Congresso Americano, depois que este teve a sua coleção original consumida em um incêndio.

Seguindo a comparação, atualmente, quase duzentos anos depois, qualquer pessoa pode ter uma coleção de trinta e oito mil livros. Essas obras, em diversos formatos digitais, são oferecidas gratuitamente no *site* do Projeto Gutenberg. Ou seja, qualquer um pode ter para si, em questão de minutos, uma coleção seis vezes maior do que aquela que Thomas Jefferson levou praticamente sessenta anos para reunir. Essa é uma mudança ímpar na história do acesso ao conhecimento e à cultura.

E pensando no impacto dessas mudanças sobre o mundo editorial, o autor de *Print is dead* afirma que é impossível que o mundo editorial não seja afetado pelas mudanças da cultura digital e se autores e editoras não abraçarem e aceitarem essas mudanças, irão perder seu público e sua receita financeira (GOMEZ, 2008).

Não estariam, além dos editores, também os bibliotecários precisando adaptar

o seu “modelo de negócio”? A resposta é sim. Os bibliotecários precisam adaptar-se aos novos suportes da informação, como sempre fizeram, desde a época dos papiros e pergaminhos.

E não há sério risco de que eles venham a perder seu “público” ou fonte de renda, principalmente se for considerado que, nessa era de abundância de informação digital, ganha proporções inéditas um dos temas clássicos da Ciência da Informação: a organização da informação. A relevância desse tema extrapola a questão do arranjo dos acervos de instituições públicas ou privadas. Ele abrange também o âmbito das coleções pessoais – uma área bastante fértil para os prestadores de consultoria na área da Biblioteconomia.

É justamente no crescente mercado da organização de coleções pessoais em mídia digital que o bibliotecário consultor pode encontrar seu melhor nicho de atuação. Contudo, para isso acontecer, ele precisa estar preparado e munido de conhecimento suficiente para criar projetos que envolvam os conceitos de bases de dados, metadados, interoperabilidade e catalogação automatizada.

1.1 IDENTIFICAÇÃO E JUSTIFICATIVA DO PROBLEMA

A despeito de serem compostas pelos mesmos tipos de itens, como livros, músicas e filmes, se comparadas às coleções institucionais, as coleções pessoais não costumam atrair muito interesse dos teóricos da Ciência da Informação. Todavia, esse é um ramo de atuação profissional bastante fértil. Por isso, os bibliotecários que atuam no setor de consultoria não restringem suas atividades às empresas, atendendo, também, a demanda por organização nos acervos pessoais.

Essa demanda, em anos recentes, ganhou um novo viés através da popularização do formato digital para armazenamento de músicas, filmes e livros. Os avanços tecnológicos baratearam os custos, tanto no que se refere à produção de conteúdo digital, quanto no que tange a sua aquisição por meio da Internet. A consequência disso é que as coleções pessoais crescem rapidamente através da constante adição de novos itens digitais.

Apesar do volume crescente das coleções pessoais de mídia digital, bibliotecários que prestam consultoria nessa área encontram dificuldade na hora de

selecionar alguma ferramenta de catalogação que seja adequada às necessidades do dono da coleção. Na literatura especializada, facilmente se encontram informações sobre programas voltados para o uso corporativo. No entanto, a adaptação de seu uso a uma escala muito menor do que aquela para a qual eles foram originalmente planejados, não compensa o custo de sua instalação.

Por outro lado, a oferta de programas, dedicados à catalogação eletrônica de coleções pessoais de mídias digitais, é grande o bastante para justificar um olhar mais atento da Ciência da Informação sobre o tema. Mesmo não sendo voltados ao uso por profissionais da biblioteconomia, mas sim por usuários domésticos, essas ferramentas implementam vários conceitos já mapeados pela área. Dentre os quais destacam-se o de “metadado” (relacionado ao registro das informações catalográficas), o de “base de dados” (relacionado à organização e recuperação das informações) e o de “interoperabilidade” (relacionado ao intercâmbio de informações entre sistemas diferentes). Partindo disso, é possível traçar o seguinte cenário:

a) as coleções pessoais de mídia digital estão cada vez maiores;

b) há uma crescente demanda pela organização desses acervos;

c) existem ferramentas que automatizam o processo de catalogação desses acervos;

d) essas ferramentas adotam determinados modelos de interoperabilidade;

e) esses modelos possibilitam a troca de metadados entre a coleção pessoal e bases de dados *on-line*.

Considerar esse cenário como o ambiente atual de atuação dos bibliotecários consultores, imbui o problema deste trabalho com uma justificativa que não é apenas teórica, mas também prática. Se esses profissionais tiverem a seu dispor o conhecimento de como aqueles conceitos se articulam na organização de acervos digitais pessoais, então, com maior facilidade poderão elaborar e desenvolver projetos nessa área.

Assim justificado o problema, resta, ainda, apresentá-lo de forma sintética.

Estando os modelos de interoperabilidade no ponto central, que interliga todos os tópicos relacionados a este assunto, optou-se pela seguinte formulação para o problema da pesquisa: os bibliotecários consultores carecem de conhecimento sobre

como se dá a interoperabilidade no processo automatizado de catalogação realizado por ferramentas voltadas às coleções pessoais de mídia digitais. É para a solução desse problema que o presente estudo tentará contribuir.

1.2 OBJETIVOS

Os objetivos desta pesquisa estão organizados em geral e em específicos, conforme serão apresentados a seguir.

1.2.1 Objetivo geral

Analisar modelos de interoperabilidade adotados por ferramentas que automatizam a catalogação de coleções pessoais de músicas, vídeos e livros digitais.

1.2.2 Objetivos específicos

Podem ser elencados como objetivos específicos:

a) localizar temas relacionados a este trabalho na literatura especializada nacional e estrangeira;

b) descrever o processo automatizado de catalogação em coleções pessoais de música, vídeos e livros;

c) descrever modelos de interoperabilidade utilizados nos processos automatizados de catalogação em coleções pessoais;

d) identificar e caracterizar as fontes utilizadas pelos processos automatizados de catalogação em coleções pessoais;

e) avaliar os modelos de representação dos metadados de acordo com o padrão internacional para descrição bibliográfica (ISBD).

2 REVISÃO DE LITERATURA

O uso de fontes de informação da Internet se destaca pela facilidade e pela velocidade de acesso. Uma necessidade informacional pode ser respondida imediatamente ou tão logo quanto o usuário for capaz de digitar os termos de busca. Os recursos presentes na Internet possibilitam, tanto ao pesquisador de alta tecnologia, quanto ao estudante de ensino fundamental, a interação com diversas formas de produção, “[. . .] sejam elas constituídas por textos, imagens, sons, fotos, vídeos, músicas, animação, multimídia, etc., que alcançam o usuário e o envolvem num espaço informacional com fins múltiplos: trabalhar, estudar, pesquisar, divertir-se, etc.” (TOMAÉL, 2008, p.vii)

Entretanto, segundo Rowley e Hartley (2008), para que o conteúdo de um recurso possa ser recuperado assim que requerido ele precisa ter sido previamente organizado. Sendo que tal organização deve respeitar três princípios da organização do conhecimento:

a) conhecimento precisa ser organizado para comunidades;

b) ao se desenvolver ferramentas de suporte a organização do conhecimento a lingüística e a estrutura cognitiva dos membros da comunidade devem ser a preocupação central;

c) padrões e padronização aumentam a interoperabilidade entre sistemas.

Em consonância com esses princípios, nesta e na próxima seção será delineada a forma pela qual a literatura especializada aborda os temas relacionados ao presente trabalho: os padrões de interoperabilidade (quanto ao modelo intercâmbio de metadados), as ferramentas de suporte a organização do conhecimento (na forma de bases de dados e programas usados para catalogação) e as comunidades alvos (no tocante à catalogação social e as coleções pessoais de mídia digital).

2.1 BASES DE DADOS

Para entender o que orienta a criação de bases de dados, explicam Rob e Coronel (2009), deve-se entender a diferença entre dado e informação. Segundo os autores, dados são fatos brutos, sem nenhum processamento que revele o seu significado. Enquanto que informação é o resultado do processamento de dados de forma que o seu significado seja revelado através de um contexto. Por exemplo, tomando-se a sequência numérica 16, 14, 12, 13, 15, 17, 16, descontextualizada não possui significado ou constitui informação. Porém, se for considerado que esses números expressam valores térmicos em graus Celsius e que se originam de medições tomadas em intervalos regulares durante o espaço de um dia em uma determinada cidade, então já se tem o contexto e com ele o significado que interliga cada dado, ou seja, tem-se informação.

Mesmo que as medições do exemplo anterior possam ser registradas, tabuladas e transformadas em gráficos usando-se apenas papel e caneta, conforme aumenta a quantidade de dados, mais trabalhoso se torna essa tarefa. Por essa razão costuma-se usar desde simples planilhas eletrônicas até complexas bases de dados computadorizadas – conforme for a quantidade de dados e variedade de contextos relacionais a serem analisados.

Neste sentido, Rob e Coronel (2009) definem base de dados como estruturas computadorizadas que armazenam uma coleção de dados brutos e metadados, através dos quais aqueles são integrados e gerenciados. De forma geral, pode se dizer que cada dado registrado em uma base deve ser acompanhado de outros dados que informem algo sobre ele, estes são os metadados. Continuando o exemplo das medições térmicas, no caso do dado bruto “16” quando for registrado na base ele pode ser acompanhado do metadado que informa de que tipo é esse valor. Ou seja, se é numérico, textual ou cronológico. Isso se torna relevante para quando for preciso manipular esse dado, pois se ele for um simples texto, não pode ser usado para cálculos, e se for numérico, então as operações matemáticas permitidas são de uma ordem, e se for cronológico, as operações são de outra.

2.1.1 Níveis de abstração

Se, por um lado, as bases de dados igualam-se por serem compostas de dados e metadados, por outro lado elas distinguem-se profundamente quanto ao modelo adotado para o arranjo desses elementos. Um modelo de base de dados nada mais é do que uma representação abstrata, relativamente simples, de uma estrutura de dados real e mais complexa (ROB; CORONEL, 2009).

Sendo cada modelo uma abstração, pode-se utilizar o padrão de níveis de abstrações, proposto pelo Standards Planning and Requirements Committee, (SPARCS) da American National Standards Institute (ANSI), para comparar e melhor entender diferentes tipos de modelos. No tocante as bases de dados, profissionais de áreas distintas possuem sobre elas diferenciadas perspectivas, conforme o grau de abstração que usam para analisá-las. Tais perspectivas podem ser enquadradas naquilo que é comumente referido como a “arquitetura ANSI/SPARC”, a qual se divide em três níveis: externo, conceitual e interno.

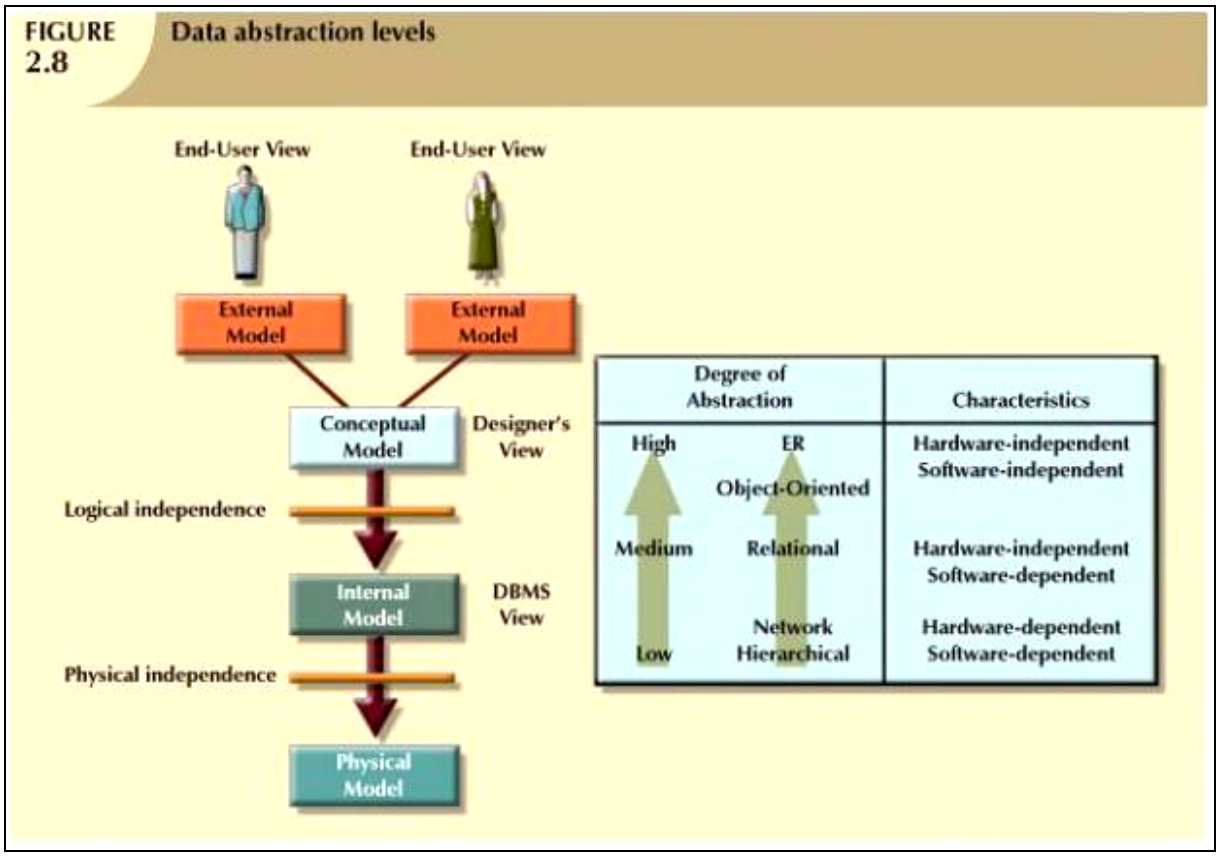
Conforme explica Date (2004), o nível externo é o nível do usuário final. É onde cada usuário pode ter uma forma diferente de interagir com a base. Um programador pode fazer consultas através de uma linguagem de programação (como Java), enquanto o outro indivíduo tem a sua disposição formulários e menus. Geralmente, esse usuário só está interessado em uma parte da base e a sua visão, provavelmente, abstrai a maioria dos dados que são importantes para o outro. Por exemplo, quem procura um livro através do catálogo de uma biblioteca, normalmente, não encontra nele a data de aquisição ou o preço dos exemplares. A visão que tem da base é uma descrição do conteúdo da obra, sua localização na estante e, talvez, a disponibilidade dela no momento. Contudo, a data de aquisição e o preço são importantes para o responsável pelo controle de aquisições e inventário da biblioteca, enquanto todas as outras informações talvez sejam dispensáveis.

Para Rob e Coronel (2009), o nível de abstração conceitual tem a função de integrar todas as visões do nível externo em uma única representação global de toda a base de dados. É neste nível que o conteúdo da informação é definido pelo desenvolvedor da base (DATE, 2004). Aqui são determinados os principais dados a serem armazenados, as chamadas “entidades” e seus “atributos” (WARD; DAFOULAS, 2006) . Voltando ao exemplo da base de dados de uma biblioteca, à

entidade “livro” é relacionada aos atributos “autor”, “título”, “editora”, “ISBN”... E o mesmo se repete com todas as entidades informacionais que a base deve abarcar. Todavia, o estabelecimento dessas relações se dá independentemente do *software* e do *hardware* onde ela será hospedada.

O nível de abstração interno representa a base de dados como “vista” pelo seu *software* de gerenciamento (como o MySQL ou o Oracle). Ou seja, neste nível o desenvolvedor precisa compatibilizar as definições apresentadas no nível superior com as restrições e características do programa no qual a base será implementada (ROB; CORONEL, 2009). Portanto, esse nível depende do *software* adotado, contudo, ainda permanece independente do *hardware*. Pois, segundo Date (2004), não inclui qualquer consideração específica sobre dispositivos físicos ou o modo como os dados digitais são neles registrados.

Figura 1 – Níveis de abstração de uma base de dados



Fonte: Rob e Coronel (2009).

Na ilustração acima é possível visualizar mais facilmente a arquitetura ANSI/SPARC e como ela é capaz de integrar os pontos de vista dos profissionais da

Tecnologia e da Ciência da Informação. Por exemplo, considerem-se as classificações que cada uma dessas áreas tem para organizar seus respectivos conhecimentos sobre bases de dados.

2.1.2 Categorização

Para um desenvolvedor de bases de dados há três tipos de modelos tradicionais:

a) hierárquico, nele a estrutura lógica é representada por uma árvore invertida, onde o seguimento superior é visto como pai daquele diretamente abaixo, sendo que cada filho só tem um pai, mas cada pai pode ter vários filhos;

b) rede, nele a estrutura lógica permite que cada seguimento tenha mais de um pai, sendo a representação mais adequada justamente a de uma rede onde cada nó pode estar conectado a vários outros;

c) relacional, nele a estrutura lógica é a de uma coleção de tabelas cujos valores das células e colunas podem ser compartilhados.

Em contrapartida, para um bibliotecário “[. . .] há dois tipos principais de bases de dados que podem estar disponíveis para usuários de informação” (ROWLEY; HARTLEY, 2008, p. 22) :

a) bases de referência, nelas o usuário visualiza informações que referem ou apontam para outro documento, organização ou indivíduo para informações adicionais;

b) bases de fontes, nelas o usuário visualiza toda a informação que precisa, sem precisar buscar informação adicionais em outro lugar, pois elas contêm a fonte dos dados originais.

Rowley e Hartley (2008) ainda subdividem cada um dos tipos principais em classificações mais específicas. Para esses autores as bases de referência podem ser dos seguintes subtipos:

a) bases de dados bibliográficas, que contém citações, referências bibliográficas ou resumos de artigos publicados;

b) bases de dados catalográficas, que apresentam os itens do acervo de uma biblioteca ou de uma rede de bibliotecas;

c) bases de dados referenciais, que listam informações cadastrais como nomes e endereços de instituições.

Enquanto que, as bases de fontes, são dividas por eles nos seguintes subtipos:

a) bases de dados numéricos, que contêm informações numéricas de vários tipos, como dados estatísticos e de levantamentos;

b) bases de dados de texto integral, que podem possuir desde notícias de jornais até artigos de periódicos científicos;

c) bases de dados textuais e numéricos, que contém materiais que combinam números e texto, como relatórios econômicos de empresas;

d) bases de dados multimídia, que contêm informações armazenadas em combinações de diferentes tipos de mídia, como sons, fotografias e vídeos.

Como cada área possui a própria classificação para bases de dados, poder-se-ia acreditar que elas seriam inconciliáveis. Contudo esse não é o caso, já que cada uma se refere a um nível de abstração diferente. Como se constata, os dois tipos principais de bases caracterizados pela Ciência da Informação contemplam apenas o ponto de vista que o usuário tem da base, o que é característico do nível de abstração externo. Por outro lado, os três tipos de modelos da Tecnologia da Informação se enquadram no nível de abstração conceitual, pois representam a base de forma global (ao contrário do nível externo) e não fazem menção à configuração do *software* (o que seria característico do nível interno).

A arquitetura ANSI/SPARC fornece fundamentos que, na medida em que elucidam e padronizam a estrutura teórica de uma base de dados, também abrem espaço para o estabelecimento de uma articulação entre a teoria adotada pelos profissionais da Tecnologia da Informação e a dos profissionais da Ciência da Informação. Esta abordagem multidisciplinar colabora na realização da meta desta seção de organizar a informação especializada sobre bases de dados relativas ao tema da presente pesquisa.

No contexto da mídia digital, o dono de uma coleção pessoal dispõe de uma cópia do documento digital. Todavia, os metadados que acompanham o arquivo eletrônico nem sempre estão completos. Por exemplo, um arquivo que contém uma determinada música pode incluir a informação sobre o seu título e não informar mais nada. Todos os outros dados, como intérprete, compositor e ano de lançamento, podem estar ausentes. Neste caso, é possível utilizar um programa de computador

para completar os metadados do arquivo importando as informações de bases de dados disponíveis na Internet.

As bases de dados consultadas por tais ferramentas de catalogação de mídia digital podem ser classificadas entre aquelas que a literatura da Biblioteconomia chama de bases de referência. Isto se justifica por posicionamentos como o de Cunha (1989, p. 46), que define esse tipo de base como aquele que apenas “ [. . .] contêm referências ou informações secundárias que identificam as várias fontes primárias”. Rowley (2002) explica esse tipo de base como sendo aquele que remete os usuários a outras fontes para que lá obtenha o documento completo. Na situação em que o usuário já tenha o documento completo, a fonte primária, a relevância desse modelo de base está justamente nos dados que fazem a remitência.

2.1.3 Metadados

Tradicionalmente metadados são definidos, segundo a etimologia do termo, como dados sobre dados. Contudo essa é uma definição por demais generalista. Escapa a ela a capacidade de explicar o que são metadados diante de um contexto específico. Nesse sentido Cox, Mulder e Tatic (2006) oferecem uma definição mais adequada, que, também, é adotada pela Society of Motion Picture and Television Engineers (SMPTE) e poderia ser adaptada para a biblioteconomia assim: a **essência** de um documento é qualquer dado necessário para representar algum tipo de experiência visual, auditiva ou sensorial; o **material** de um documento é composto por qualquer combinação de essência e dados; o **metadado** é qualquer dado que carregue informação sobre o documento material; enquanto que o **conteúdo** é formando pela combinação do material com os seus metadados. Seguindo esse conjunto de conceitos pode-se contemplar o seguinte exemplo: as palavras que compõem um livro são sua essência; as palavras fixadas em um suporte (sejam páginas físicas ou sinais digitais) materializam o livro; as informações sobre o livro (quantidade de páginas ou tamanho do arquivo) são seus metadados; e a biblioteca que guarda o livro e tem registro de seus metadados possui o conteúdo dessa obra.

Desse modo, pode-se perceber uma das vantagens dessa abordagem ao

conceito de metadados para os bibliotecários. Eles deixam de ser algo oriundo de um contexto pouco familiar, no qual o “[. . .] crescimento acelerado e caótico da Web coloca na ordem do dia o problema de identificar, recuperar e avaliar a infinidade de recursos os mais diversos” (MARCONDES, 2006, p.95). Os metadados voltam a ser algo mais simples com o qual os bibliotecários já estão familiarizados há séculos; os registros catalográficos (CASTRO; SANTOS, 2007).

Ainda que o significado do termo possa se alterar de uma comunidade para outra, a definição da National Information Standards Organization (NISO) também se aproxima do conceito do registro em um catálogo. Segundo a NISO (2004, p.1), “[. . .] metadados são informações estruturadas que descrevem, explicam, localizam, ou tornam mais fácil recuperar, usar ou gerenciar um recurso de informação” e podem ser classificados conforme três finalidades distintas:

a) metadados descritivos, atendem a necessidade de identificação e de descoberta de um recurso, podem incluir elementos como título, autor, resumo e palavras-chave;

b) metadados estruturais, indicam como os componentes do recurso são montados, por exemplo, indicando a ordem de páginas e capítulos ou a disposição de imagens;

c) metadados administrativos, fornecem informações sobre como gerenciar um recurso, podendo abranger desde a data de criação até o controle de acesso e uso.

2.2 MODELOS DE INTEROPERABILIDADE

A Internet nada mais é do que uma rede que interconecta computadores e as aplicações que neles são executadas. A interoperabilidade é o conceito que explica a habilidade de interconexão de computadores diferentes através de aplicações também distintas. Isto é relevante para a Tecnologia da Informação, na forma destacada por Rowley (2002, p.61):

> A eficácia potencial da tecnologia da informação depende

> grandemente da eficácia das redes de telecomunicações. Uma das principais finalidades de uma rede é oferecer compatibilidade ponto-a-ponto, de modo que qualquer terminal ou sistema possa se conectar com qualquer outro.

Parafraseando o ponto de Rowley, a eficiência atual da própria Ciência da Informação depende grandemente da eficiência da interoperabilidade. Pois uma das principais finalidades da interoperabilidade está relacionada a “[. . .] realizar uma busca através de conjuntos diversos de metadados registrados e obter resultados significativos” (CAPLAN, 2011, p. 33).

No ambiente da Internet, os desenvolvedores de aplicações se preocupam com vários tipos de interoperabilidade. Segundo Caplan (2011) o conceito de interoperabilidade pode significar:

a) que duas aplicações compartilham do mesmo protocolo de comunicação;

b) que um cliente pode interagir com vários servidores, ou;

c) que dados podem ser reusados em diferentes contextos.

Pode-se considerar que, no que tange aos aspectos técnicos, a unificação desses três significados cristaliza-se em um conceito mais restrito, o de **computação distribuída**.

Um sistema que adote alguns padrões técnicos pode alcançar a um ou a outro dos significados de interoperabilidade apontados por Caplan. Entretanto, segundo Arms (2000), os custos desse projeto podem ser um sério impeditivo. Principalmente se for levado em conta que leva tempo modificar um sistema para que ele suporte um novo padrão, e quando a alteração estiver concluída, talvez já exista uma nova versão do padrão ou a comunidade de usuários já pode ter decidido seguir outro caminho.

Dependendo da realidade de cada biblioteca digital, a abordagem à interoperabilidade pode variar. Por exemplo, examine-se o caso das bibliotecas digitais criadas em universidades e institutos de pesquisa visando à divulgação de material acadêmico. Segundo Kim, Choo e Chen (2010), atentar aos padrões de interoperabilidade é fundamental para uma melhor acessibilidade, organização, recuperação, visibilidade e disseminação das informações acadêmicas. Ou seja, a interoperabilidade é um meio de tornar os recursos digitais mais fáceis de localizar,

explorar e utilizar por pesquisadores que, do contrário, estariam perdidos em um crescente e, por vezes, caótico mar de informações acadêmicas. O principal desafio para se atingir a interoperabilidade entre coleções de bibliotecas digitais é a questão da descentralização e heterogeneidade das unidades detentoras dessas coleções.

Um objetivo aparentemente simples, como conseguir que dois diferentes repositórios de teses possam ser pesquisados através da mesma *interface,* envolve um estudo das linguagens de consulta, protocolos de recuperação de informação e capacidades de processamento de cada um deles. Os custos de implementação e manutenção dessa interface de busca não variam só pela quantidade de repositórios a serem integrados, mas também, e principalmente, em função da tecnologia escolhida para proporcionar essa interoperabilidade.

Isso significa que é válido adotar uma tecnologia mais simples (uma interoperabilidade mais limitada em funcionalidades), mas com custos que permitam uma aplicação mais ampla. De modo geral, Arms (2000) aponta três distintas abordagens a esse problema:

a) poucas funcionalidades e custo baixo, como os padrões da Web (HTTP, URL e HTML) que pela simplicidade tiveram ampla adoção, mas permitem funções mais complexas;

b) muitas funcionalidades e custo alto, como os padrões Z39.50 e SGML que só são populares em círculos restritos, que valorizam suas funcionalidades, pois seus custos dificultam uma ampla adoção por outras comunidades;

c) importantes funcionalidades e custo moderado, como os padrões Dublin Core e XML que se destacam pela opcionalidade dos campos e pela relação próxima com os padrões da Web.

Sendo que, atualmente, a interoperabilidade não visa unicamente comunicar um computador ao outro, mas esses a toda sorte de dispositivos móveis (como telefones celulares e *tablets*), então o papel da Biblioteconomia pode precisar ser ampliado. Para Mucheroni e Silva (2011, p. 16):

> A Ciência da Informação tem longa tradição e experiência na descrição de conteúdos; e a crescente heterogeneidade de dispositivos e equipamentos pode ser auxiliada por esta descrição desenvolvida em diferentes padrões de metadados e linguagens de marcação, colaborando com a interoperabilidade nestes ambientes.

É possível que essa seja uma abordagem que pareça distante do cotidiano da maioria dos bibliotecários. Contudo, é preciso manter em mente a função desempenhada pelo profissional da Ciência da Informação, por exemplo, na composição de bases de dados de referência e em como a maioria dos novos modelos de *smartphones* já consegue consultá-las.

2.2.1 Busca federada de informações

A busca federada, conforme Kim, Choo e Chen (2010), – mesmo através de nomes diferentes, como: pesquisa multi-base de dados (*multi-database search*), meta-busca (*meta-search*) ou pesquisa distribuída (*distributed search*) – foi estudada e discutida por anos como a forma principal de se realizar buscas através de várias fontes dispersas e autônomas. Contudo, os autores a descrevem como uma interface comum de busca unificada que recupera dados dispersos realizando a busca no respectivo lugar que os armazena. A consulta do usuário nessa interface é repassada para cada base de dados em uma linguagem de consulta que ela possa entender, os resultados são então normalizados, produzindo visão unificada dos dados para o usuário. Assim, o usuário só precisa aprender a usar uma única interface de pesquisa para conseguir recuperar, simultaneamente, resultados de diferentes fontes de dados heterogêneas.

Outras vantagens do modelo de busca federada, citadas Kim, Choo e Chen (2010), são:

a) dispensa o armazenamento centralizado, o que reduz o custo de *hardware*;

b) precisa apenas de uma aplicação de baixo custo de processamento para acessar um grande número de repositórios;

c) atualiza os resultados a cada nova busca, sendo esses os dados mais novos de que dispõem os repositórios alvos.

Também pode ser incluído nessa lista de vantagens o fato de a maioria dos projetos de busca federada utilizarem o protocolo Z39.50. Segundo Simeoni (2004), as especificações do uso do Z39.50 na descoberta de recursos informacionais já

foram padronizadas e amplamente testadas através de várias implementações pela comunidade biblioteconômica. Todavia, nem mesmo o uso desse protocolo evita que flutuações na velocidade da rede provoquem lentidão ou indisponibilizem a comunicação com alguma fonte de dados mais distante. Isso impacta diretamente na continuidade, confiabilidade, agilidade e eficácia da prestação do serviço proposto pela busca federada.

Para que a qualidade e agilidade do serviço prestado por uma busca federada estejam a altura do esperado por seus usuários, que a comparam com os populares motores de busca da Internet, os custos de implementação tornam-se quase imensuráveis (SIMEONI, 2004). E esses custos se ampliam sempre que uma nova unidade autônoma, na gestão de sua própria base, é agregada à federação de busca. Por esse motivo o modelo de colheita de dados tornou-se tão popular.

*2.2.2 Colheita (*harvesting*) de metadados*

O modelo de interoperabilidade por colheita (*harvesting*) de metadados se diferencia da busca federada por primeiro reunir os dados em uma central de armazenamento, a partir da qual eles serão consultados pelos usuários. De acordo com Kim, Choo e Chen (2010), essa abordagem fornece melhor performance, uniformidade e consistência se comparada à busca federada, o que, em contrapartida, significa um maior custo de manutenção e armazenamento dos dados coletados.

Para padronizar o uso desse modelo no intercâmbio de informações acadêmicas, a Open Archives Initiative (OAI), explica Suleman (2006), criou o Protocol for Metadata Harvesting (PMH). Esse protocolo atende a necessidade de diminuir as barreiras das redes de computador, representadas pela largura de banda e tempo de resposta, através da implementação da transferência de metadados em uma base periódica (diária, semanal ou mensal). O proprietário dos metadados é referido pela OAI como o **provedor de dados** (data provider), enquanto que o fornecedor do serviço de busca é referido como o **provedor de serviços** (service provider). Desse modo, a colheita é definida como o ato de transferir os metadados

do provedor de dados para o provedor de serviços através de uma ferramenta de *software* chamada de colheitadeira (*harvester*) (SULEMAN, 2006).

Segundo Breeding (2002), o PMH trabalha com metadados de uma forma bastante flexível. A escolha do formato padrão para o registro dos metadados a serem transferidos depende unicamente de um acordo entre os provedores de dados e os provedores do serviço. Outra característica, apontada por Breeding (2002), é o fato dele aproveitar o protocolo Hypertext Transfer Protocol (HTTP), o mesmo usado pelas páginas da Web, para fazer a transferência dos dados.

O PMH especifica um conjunto de “requisições” e “respostas” que pode ser definidos como segue:

a) GetRecord: requisita a transferência de um registro de metadados particular;

b) Identify: requisita ao provedor de dados para enviar as informações de sua própria identificação;

c) ListIdentifiers: requisita que ao provedor de dados que forneça a lista de cabeçalhos de registros que satisfazem a uma busca;

d) ListMetadataFormats: requisita ao provedor de dados que informe a lista de formatos de metadados suportados;

e) ListRecords: requisita a lista de registros contidos no repositório;

f) ListSets: requisita a lista de conjuntos nos quais foi dividida a coleção do repositório.

Essas requisições permitem que a colheitadeira interaja com o provedor de dados, mesmo que seja em um nível de interatividade muito menor do que um usuário estaria acostumado a encontrar em um *site* de busca na Internet. Isso porque, segundo Breeding (2002), o único objetivo de uma seção de colheita é transferir em massa todos os registros novos e modificados desde a seção inicial. Suleman (2006) detalha da seguinte forma o funcionamento do processo de colheita:

a) o provedor de serviços aciona sua colheitadeira para colher os metadados de um provedor de dados. Se os metadados ainda não tinha sido colhidos antes, a colheitadeira requisita todos os metadados em um formato especificado;

b) o provedor de dados responde com tantos metadados quanto ele for capaz de processar e envia junto para a colhedora um sinal de retomada;

c) a colheitadeira repassa os registros ao prestador de serviços para que sejam inseridos no sistema de busca;

d) se a colheitadeira encontra um símbolo de retomada no final do fluxo de registro, ela envia uma solicitação subsequente ao provedor de dados com este sinal como um parâmetro;

e) o provedor de dados responde, então, com mais um pedaço de sua coleção de registros e um novo sinal de retomado, se for preciso. Este processo continua se repetindo até que todos os registros forem transferidos;

f) quando todos os registros foram transferidos, a colhedora encerra suas atividades;

g) em seguida, periodicamente, o provedor de serviços aciona a colheitadeira para obter os registros alterados desde a última colheita.

A Internet causou uma revolução na acessibilidade da informação e a adoção de metadados representa uma melhora significativa na forma como a informação pode ser colhida (HIRWADE; BHERWANI, 2009). Todavia, como visto, nesta seção, a colheita de informação é um processo que ocorre na interface máquina a máquina, e que depende da adoção de um mesmo padrão para o protocolo de transferência e de um acordo quanto ao formato de registro dos metadados pelas partes envolvidas. Esses são os pontos principais que em que a colheita se distingue da raspagem de dados.

*2.2.3 Raspagem (*scraping*) de dados*

A maior parte dos *sites* da Internet, conforme Rhodes e Goerzen (2010), são desenvolvidos primordialmente para os olhos humanos. São poucos os *sites* que fornecem aos seus usuários uma Application Programming Interface (API), a qual permita uma interação automatizada via interface máquina a máquina. Isso significa

que tarefas repetitivas como, por exemplo, copiar fotos distribuídas através de várias páginas de um site, precisam ser executadas através da interface humano-máquina. No entanto, isso não significa que a tarefa de obter dados através dessa interface precise ser feita manualmente.

Para McGregor (2012), o uso de programas de computador para automatizar a transferência, a partir da Web, de grandes quantidades de dados que dificilmente poderiam ser obtidos de forma manual, caracteriza a técnica chamada de raspagem de Web (*web scraping*).

O resultado da raspagem de dados da Web não se mostra apenas em *sites* que combinam informações vindas de outros, como os de comparação de preço e os agregadores de notícias, mas também no próprio trabalho de jornalistas. Ela pode ser uma poderosa ferramenta que, segundo McGregor (2012), permite aos jornalistas combinar registros eletrônicos para encontrar histórias que não poderiam ser facilmente descobertas de outra forma. As agências governamentais estão cada vez mais colocando dados *on-line*. Normalmente, os dados são escondidos atrás de uma interface Web que só permite pesquisas por termos determinados. O fato de agências e organizações, nem sempre liberarem suas informações na forma de bases de dados bem formatadas, talvez seja o principal motivo para jornalistas e pesquisadores aprenderem a programar (NGUYEN, 2011).

Se por um lado, uma página da Web pode ser a fonte dos dados mais precisos e atualizados, por outro lado, alertam Alba, Bhagwan e Grandison (2008), a raspagem desses dados pode falhar ou até capturar dados errados se houver uma alteração no código fonte do *site* de origem. Dependendo de como foi criado o programa de raspagem, uma modificação grande, como a troca do layout do site, pode não afetar a captura dos dados, mas uma modificação menor, como a correção de um erro de digitação, pode fazer com que ele não consiga mais recuperar dado algum.

Por esse motivo, afirmam Alba, Bhagwan e Grandison (2008), é preferível usar uma API, quando houver uma disponível, já que ela atende justamente a necessidade de fornecer um acesso automatizável, máquina a máquina, que não é afetado por mudanças na interface humana do *website*. E, através da API de um *site*, pode-se recuperar dados específicos, da mesma forma que a raspagem. Além disso, em alguns casos, conseguem-se encontrar dados que não estão disponíveis

para o usuário a partir da interface padrão. Apesar dessas vantagens, ponderam os autores, APIs nem sempre apresentam um comportamento confiável e consistente, o que é agravado pela pouca documentação e até mesmo funcionalidades estranhas, tornando, muitas vezes, difícil tirar o máximo proveito dessas ferramentas.

Dessa forma, mesmo que se recorra ao uso de uma API, o modelo de interoperabilidade baseado em raspagem de dados da Web diferencia-se do modelo de colheita e do modelo de busca federada, por não contemplar uma padronização negociada e acordada entre as partes. Esse fato reduz consideravelmente os custos de implementação da raspagem de dados, quando comparada a outros dois modelos. Isso porque não demanda da fonte de dados que ela adote mais um protocolo de comunicação (como o Z39.50 para a busca federada), nem sugere que ela modifique a apresentação de seus dados para outro formato (como o Dublin Core para a colheita).

2.3 PRODUÇÃO DE METADADOS

Se cada tipo de metadado atende a uma finalidade diferente, é importante levantar a questão: não seria mais conveniente que cada tipo de metadado tenha seu próprio criador? Por exemplo, no caso de uma coletânea de artigos em formato digital a ser incluída no catálogo de uma biblioteca, pode-se imaginar que os metadados descritivos (título, autor e palavras-chave) sejam fornecidos pelo editor da coletânea; os metadados estruturais sejam formatados pelo bibliotecário (incluindo analíticas de cada artigo no sistema); e os metadados administrativos sejam escolhidos pelo gerente da instituição mantenedora da biblioteca (bloqueando o acesso a alguns usuários e liberando para outros). Contudo, a NISO (2004) alerta que nesse tipo de cenário podem surgir problemas de qualidade dos registros, principalmente se a responsabilidade sobre a criação dos metadados for compartilhada com pessoas não familiarizadas com catalogação, indexação ou controle de vocabulário.

Preocupada com a avaliação da qualidade dos registros, a NISO (2004) arrola seis princípios aplicáveis aos bons metadados:

a) devem ser adequados aos materiais da coleção, aos usuários da coleção, e prever o uso corrente e futuro do objeto digital;

b) devem suportar interoperabilidade;

c) devem usar vocabulários controlados padrões para refletir o “o quê”, “onde”, “quando” e “quem” do conteúdo;

d) devem incluir uma declaração clara sobre as condições e termos de uso do objeto digital;

e) devem ter as qualidades de arquivabilidade, de persistência, de identificação única (etc.), mantendo-se confiáveis (*authoritative*) e verificáveis;

f) devem suportar a gestão a longo prazo de objetos em coleções.

Esses princípios não são só aplicáveis aos metadados na maneira como são utilizados em bibliotecas. Atualmente profissionais de várias áreas passaram a se preocupar com padrões e métodos de descrição de informação. Um exemplo é apresentado pelo caso das redes de televisão, que, segundo Cox, Mulder e Tatic (2006), devem se preocupar com a precisão dos seus metadados pois eles podem ser uma importante fonte de receita, permitindo reempacotar seriados para o consumidor doméstico, reaproveitar gravações para uso interno e licenciar gravações para o usuários externos.

No caso das bibliotecas, a preocupação com a precisão dos metadados do catálogo não é algo novo, mas perceber que eles podem representar uma fonte de receita, talvez seja. Um catálogo *on-line* confiável pode atrair novos usuários e um número maior de usuários pode alavancar uma positiva negociação sobre verbas junto à instituição mantenedora. Essa mesma linha de raciocínio embasa o desenvolvimento do plano de negócio das bases de dados *on-line* sobre livros, músicas, e vídeos – principalmente se além dos metadados, elas também fornecerem o conteúdo.

2.3.1 Catalogação tradicional

O modo tradicional pelo qual são incluídos novos itens nos catálogos das bibliotecas envolve o esforço de um bibliotecário em identificar as informações pertinentes no item e copiá-las, de forma padronizada, para o suporte do catálogo,

seja ele eletrônico ou não. Todavia, em se tratando de um catálogo eletrônico, esse esforço de elaboração dos metadados não precisa ser unicamente do bibliotecário. Em seu texto de 2004, a National Information Standards Organization (NISO), compartilha e flexibiliza essa responsabilidade na sua resposta para a pergunta: quem cria os metadados?

Para a NISO, a responsabilidade sobre criar os metadados está com a equipe técnica que inicialmente digitalizou ou criou o objeto digital; podendo, até mesmo, ser os metadados gerados através de um processo automatizado. Em um momento seguinte, pode-se recorrer ao originador do recurso, talvez um pesquisador, para criar um “esqueleto” com os elementos que mais prontamente puderem ser fornecidos, sendo o resultado disso revisado e complementado com o auxílio de indexadores ou outros profissionais da informação. Ou seja, a resposta que a NISO oferece para “quem cria os metadados?” varia conforme as possibilidades de cada projeto ou disciplina, mas costuma envolver um esforço coletivo.

Apesar do contexto parecer mais moderno e haver novas ferramentas envolvidas no processo, essa ainda é uma abordagem bastante tradicional à catalogação. Conforme Gardner (2012), a moderna catalogação, com seus mais de 100 anos, ainda consiste no ato dos catalogadores de criar registros substitutos para os itens da biblioteca, ou acessados através do catálogo dela. Esses substitutos já foram fichas catalográficas guardadas em um fichário de aço. Hoje, nas bibliotecas informatizadas, esse substituto do item está guardado em uma base de dados, mas ainda contém uma descrição, um número de classificação, os termos validados no controle de vocabulário e autoridades. Tudo isso para ajudar com o agrupamento e localização de itens semelhantes.

2.3.2 Catalogação social

A catalogação social subverte, em parte, essa tradição. Segundo seus princípios, a participação de profissionais na geração dos metadados não é requerida, nem os substitutos do item catalogado estão restritos a uma única descrição, uma classificação ou a um único vocabulário controlado.

O termo catalogação social, segundo Giustini, Hooker e Cho (2009), refere-se

a aplicações baseadas na Web que permitem aos seus usuários localizar e registrar seus livros e outros itens em um inventário *on-line* ou sistema de arquivamento pessoal. Para Spiteri (2009), a catalogação social, como um fenômeno que se desenvolve na Internet, surgiu na esteira do aumento de popularidade das redes sociais.

Nesses sites, que se convertem em verdadeiras bases de dados (como as vistas na subseção 2.1), os próprios usuários catalogam os itens que eles possuem (CDs de músicas, DVDs de filmes e todo tipo de livros). Mas, mais do que isso, nesses sites os usuários das bases podem partilhar publicamente os catalogados de suas coleções pessoais, além postar opiniões sobre itens catalogados por outros, através de sistemas de grupos de discussão, de etiquetagem (*tagging*) e de classificação. Em outras palavras, como colocou Spiteri, esses sites servem como um catálogo compartilhado e interativo, desenvolvido pelo próprio usuário.

Essas características são sintetizadas por Giustini, Hooker e Cho (2009) nas ações de criar e compartilhar bibliografias, construir novas comunidades digitais e reforçar a encontrabilidade. Essas características estão interligadas. Uma maior capacidade de compartilhar resenhas e comentários, através de uma comunidade com membros de todo o mundo, amplia e apura o detalhamento de informações bibliográficas, o que produz catálogos com resultados de busca melhores, mais precisos e relevantes. Para os autores, esse tipo de processo – no qual grandes grupos de usuários amadores elaboram um produto, cuja qualidade final teria um custo proibitivo, caso tivesse sido empregado mão-de-obra profissional – está relacionado ao conceito de “*crowdsourcing*”.

A palavra em língua inglesa para terceirização é *outsoursing*, cujo significado está ligado à contratação de uma fonte (*source*) externa (*out*) de mão-de-obra para realização de uma função anteriormente atribuída a profissionais da própria empresa. Por associação surgiu a palavra *crowdsourcing* para designar a situação em que a fonte da mão-de-obra é uma multidão (crowd) de usuários reunidos através da Internet. A motivação do aumento de popularidade deste processo é explicada por Howe (2009) pelo fato de que, dadas as condições corretas, a multidão sempre terá uma performance melhor que qualquer número de empregados. Para esse autor, há dois atributos comuns a todo projeto de *crowdsourcing*:

a) a motivação primária dos participantes não é o dinheiro;

b) e os participantes doam seu tempo livre para a causa.

Spiteri (2011) elenca as três fontes primárias de motivação para alguém utilizar *sites* de catalogação social:

a) organizar coleções pessoais;

b) auxiliar na localização de itens de interesse;

c) comunicar itens de interesse aos outros.

Todavia, neste ponto surge a dúvida: como um projeto no qual os envolvidos não são pagos e trabalham unicamente nas suas horas de folga pode funcionar? Howe (2009) aponta duas justificativas. A primeira, afirma que o *crowdsourcing* tem a capacidade de formar uma meritocracia perfeita, na qual não importa a raça, o gênero, a idade ou a qualificação profissional, nela só importa a qualidade do trabalho desempenhado. A segunda, afirma que o *crowdsourcing* tem uma espantosa tendência para atrair pessoas capacitadas vindas dos mais improváveis lugares, como se fosse um imenso mecanismo de localização de talentos.

Contudo, não é todo o projeto que tem a sorte de angariar *on-line* o talento e o conhecimento esperado. Giustini, Hooker e Cho (2009) comentam a estrutura informal e a inconsistência da linguagem usada em *sites* de catalogação social, problemas esperados em se tratando de um ambiente onde a indexação não é moderada. Os problemas mais comuns apontados por esses autores foram a falta de uma hierarquia entre os termos (como a presente em um tesauro) e a falta de uma padronização da terminologia (através de um vocabulário controlado), e isso prejudica a qualidade da indexação desse tipo de *site*.

A supervisão de um profissional da Ciência da Informação poderia ajudar a mitigar toda sorte de ambiguidades conceituais indesejada, sejam elas fruto de um sistema de *tags* mal ajustado ou de um formulário de catalogação mal organizado. Todavia, um contraponto a essa visão é apresentado por Spiteri (2011), o qual afirma que embora o conteúdo dos metadados seja pobre se comparado com os de um catálogo de biblioteca, os sites de catalogação social compensam no conteúdo que é fruto da interação social na qual os interesses de leitura são compartilhados.

Talvez um catálogo construído pelo usuário e para o usuário não deva ser avaliado do ponto de vista de um bibliotecário, mas sim da óptica de um cientista da informação preocupado em saber se determinadas necessidades informacionais estão sendo atendidas a contento.

2.4 COLEÇÕES PESSOAIS DE MÍDIA DIGITAL

O processo de evolução tecnológica causou a diminuição da popularidade dos cadernos que serviam de diários pessoais e o aumento da popularidade dos álbuns de fotografia – como forma de registro da trajetória pessoal de um indivíduo. Na mais recente mudança de tecnologia, as coleções pessoais deixaram o papel e os formatos analógicos rumo a versão digital. Essa mudança pode ser justificada pelo fato de que "[. . .] ter a coleção de fotos, vídeos e músicas disponíveis em forma digital permite que se faça muitas coisas que eram difíceis ou impossíveis de se fazer antes, nos tempos analógicos" (BALDRIGE, 2009, p.9). Facilidade de uso e novas possibilidades, em geral, são as chaves da atração de toda nova tecnologia e esse é justamente o caso da **mídia digital**.

O termo "mídia digital" é aqui empregado não no sentido do suporte onde são gravados os sinais eletrônicos (como CDs e DVDs), mas sim no sentido do conteúdo que tais sinais representam, como, por exemplo, a música, o vídeo ou o livro cuja essência está apresentada em forma digital. Esse uso do termo coaduna com a definição de mídia digital proposta pela Digital Media Alliance Florida, citada por Pennycook (2011, p. 188): "The creative convergence of digital arts, science, technology and business for human expression, communication, social interaction and education." Essa, sem dúvida, é uma definição ampla o bastante para englobar os três tipos de mídias digitais abordados no presente estudo: músicas, vídeos e livros.

Através da Tecnologia da Informação – representada por dispositivos computadorizados (como *desktops*, *tablets* e *smartphones*) e combinada à mídia digital – novas facilidades e possibilidades embrenharam-se profundamente na vida de muitas pessoas. Por isso, as coleções pessoais de mídia digital são compostas de itens muito mais variados do que apenas músicas, vídeos e livros. Beagrie (2005) descreve a variedade dos itens das coleções pessoais digitais, assim:

Combinando as posições de Beagrie (2005) e Lee (2011) essas duas declarações pode-se arrolar as seguintes características de uma coleção pessoal:

a) é um agregado da trajetória pessoal de um individuo;

b) é composta de materiais originados da vida privada, do trabalho, da educação, de fontes de conteúdo e comunidades externas;

c) é destinada ao uso e referência privados, mas pode ser compartilhada com outros indivíduos do trabalho, da família ou de grupos de interesse;

d) é diverso e complexo o seu direito de posse e de propriedade intelectual;

e) é gerida ao longo do tempo como uma unidade relativamente coerente;

f) reflete algo importante sobre o indivíduo.

Neste ponto, é possível relacionar as coleções pessoais em mídia digital a uma espécie de complemento à própria memória do indivíduo. Isso remete o assunto a um dispositivo de uso individual que seria algo como uma biblioteca ou arquivo privado, chamado memex, imaginado por Bush (2001), um dispositivo no qual um indivíduo armazene todos os seus livros, registros e comunicações e que possa ser rápida e flexivelmente pesquisado.

No texto, originalmente publicado em 1945, pode-se vislumbrar não só aquilo que mais tarde veio a ser chamado de computador pessoal, mas também uma de suas mais populares funções na atualidade: armazenar todos os livros, registros e comunicações de um indivíduo, mas, mais especificamente, todas as fotografias, músicas e vídeos. Veith (2006) concorda com alguns pontos de semelhança entre o memex e o computador pessoal, principalmente no que tange as coleções de material digital e digitalizado que cabem em um *desktop*, mas afirma que esse ainda não é o tipo de ambiente unificado retratado no artigo de Bush (2001). São dois os pontos de distinção levantados por Veith (2006):

a) a maioria do material não entra no sistema através de uma forma única (embora muito se adquiria através da Web);

b) encontrar itens em um computador pessoal, com muita capacidade de armazenamento, muitas vezes é mais lento do que encontrá-los na Internet.

Há dois motivos que justificam ser mais rápido para um usuário localizar na Internet itens que ele já tem em seu computador. O primeiro é indicado por Lee (2011), segundo ele, há vários fatores que tornam menos provável que um indivíduo apague arquivos digitais, dentre eles:

a) a exclusão demanda tempo, atenção e a certeza de que o item não será mais necessário;

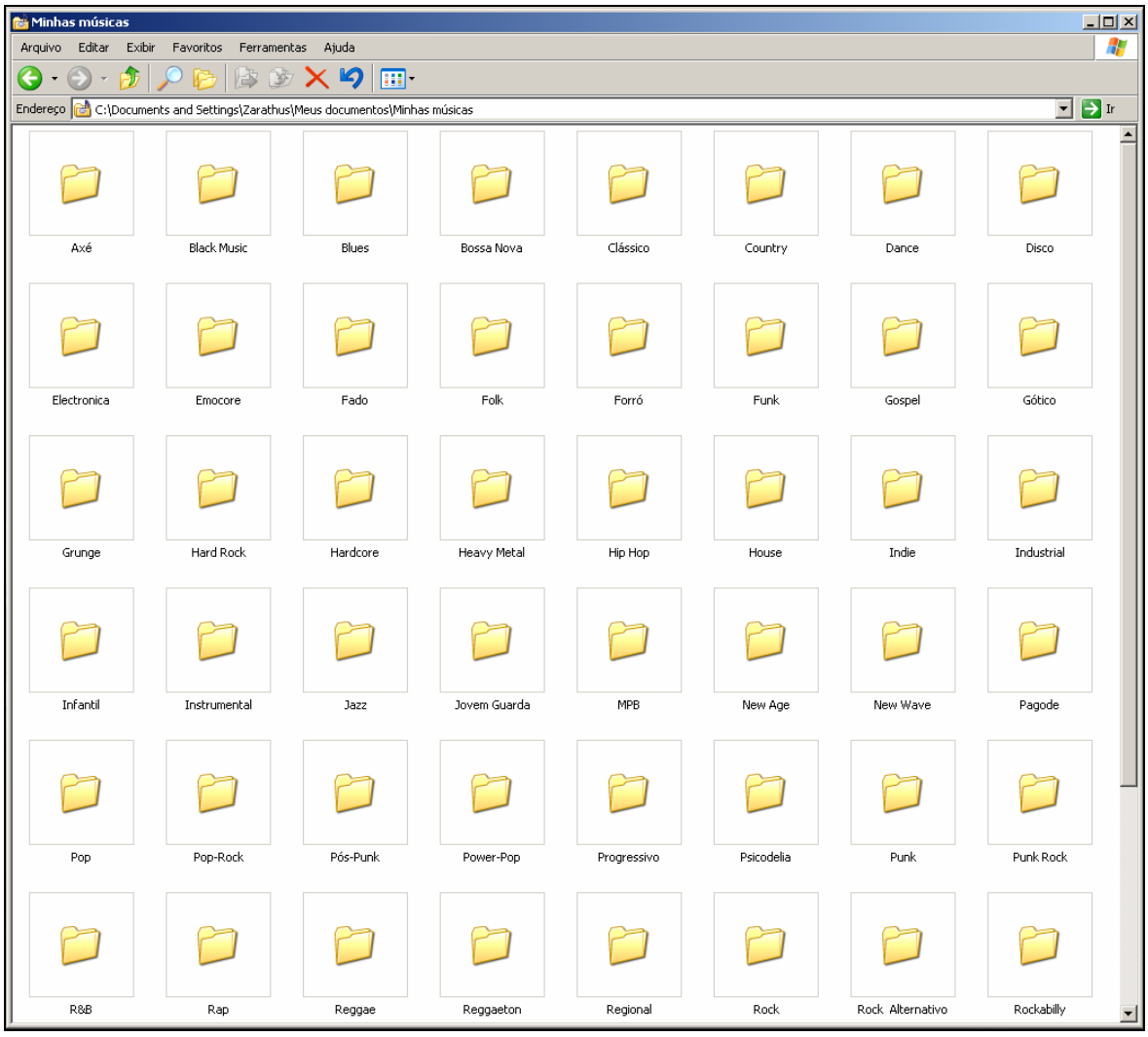
b) o custo do meio de armazenamento diminui a cada ano, o que significa que muitas pessoas não precisam se preocupar com a falta de espaço;

c) o número de dispositivos de armazenamento cresce a cada ano, um indivíduo pode ter disposição, por exemplo: *pen drives*, discos externos,

câmeras, telefones celulares além do próprio computador.

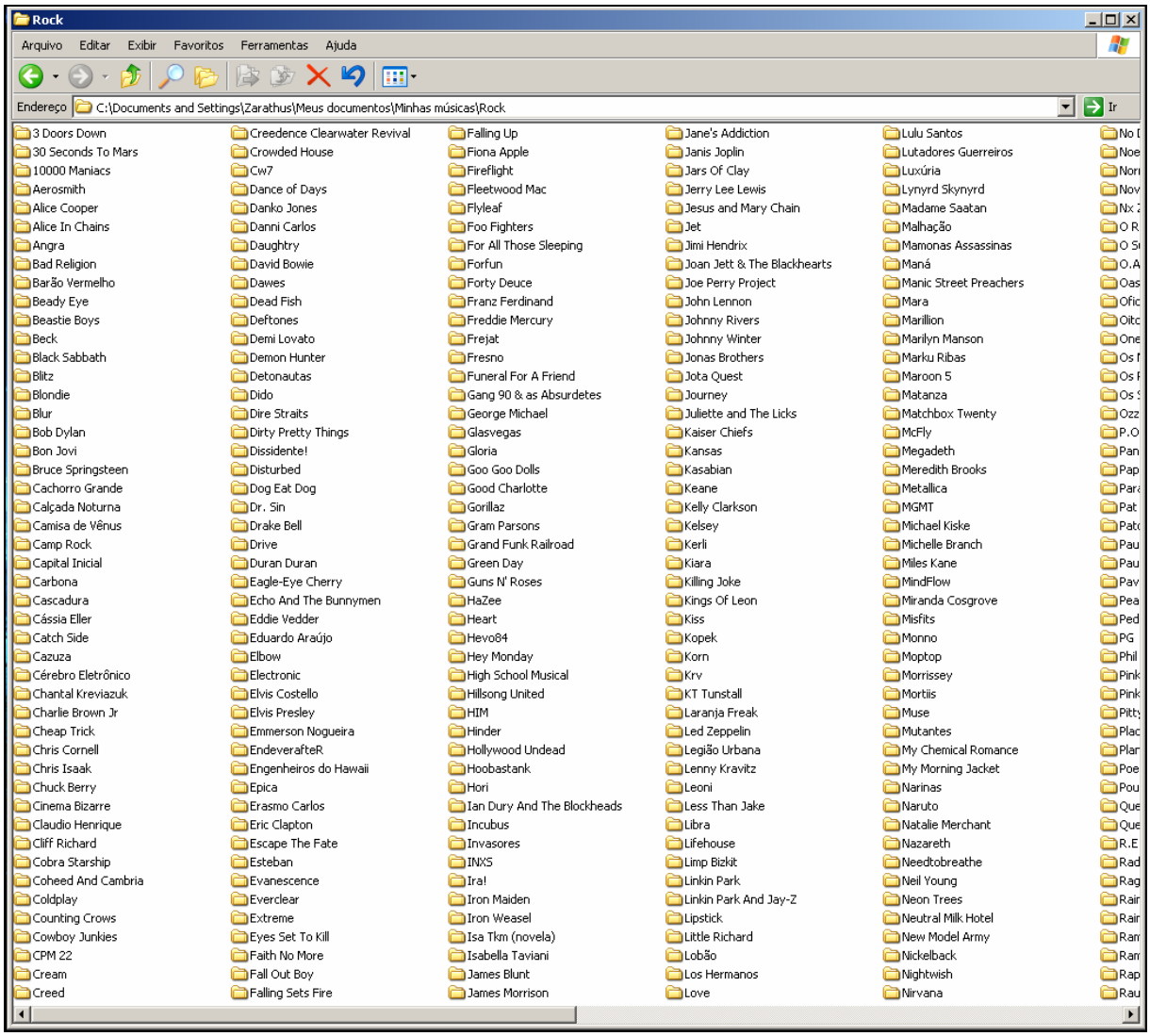
O segundo motivo que faz com que seja mais lento encontrar arquivo em um computador pessoal, para Veith (2006), é a própria organização das coleções em pastas e subpastas. Para boa parte dos usuários de computador, organizar suas coleções de músicas, vídeos e livros digitais significa, tão somente, agrupá-los nas pastas do sistema operacional correspondentes ao tipo de mídia. Todavia, quanto maior o tamanho e mais variada for a coleção, maior será a dificuldade da busca. Mesmo que se utilize vários graus de subpastas, classificando, por exemplo, as músicas, em pastas para os gêneros dentro das quais os arquivos são organizados por artistas e álbuns, chega-se a um ponto em que se percebe que essa é apenas uma solução paliativa.

Figura 2 – Pasta de músicas classificadas em vários gêneros musicais



Fonte: autor

Figura 3 – Subpasta do gênero Rock com músicas classificadas por artistas



Fonte: autor

No cenário ilustrado acima, torna-se bastante difícil localizar uma música sem saber qual o gênero no qual foi classificada e o artista que a interpreta. Caso o dono da coleção saiba apenas o nome da música, ele poderia tentar usar a ferramenta de pesquisa de arquivos e pastas nativa do sistema operacional. Contudo, essa estratégia não funcionará se o nome do arquivo não coincidir com o nome da música que ele contém. Isso demonstra outros dois pontos levantados por Veith (2006):

a) coleções de informação tornaram-se muito mais complexas no *desktop* do que foram descritas no memex de 1945;

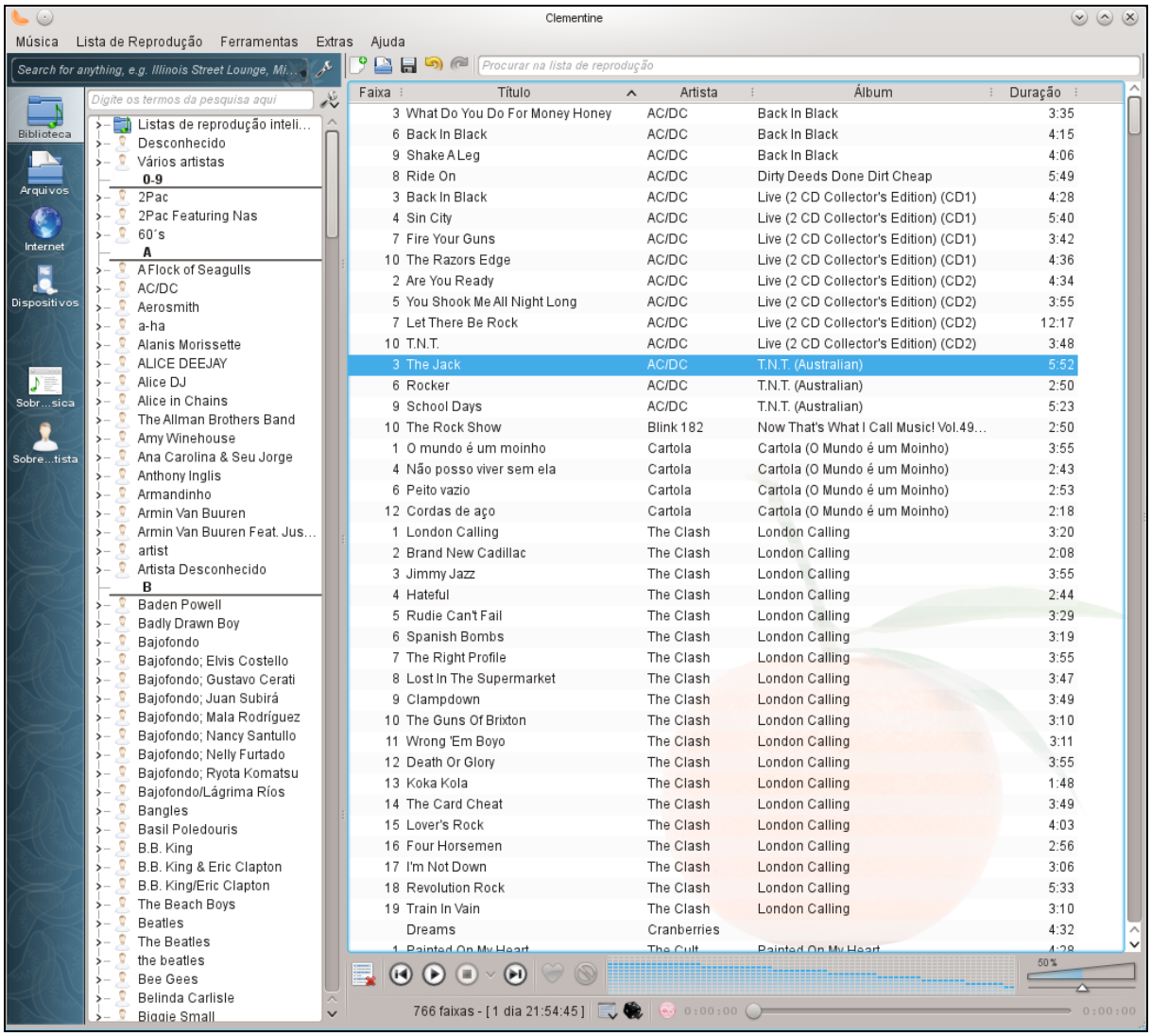
b) as ferramentas necessárias para lidar com a busca de informações dentro

desta memória estendida precisam ser similarmente muito mais complexas;

c) em um ambiente unificado, pode haver melhores formas de encontrar as coisas do que procurar através de pastas digitais ou usar ferramentas específicas de pesquisa.

Os desenvolvedores dos programas responsáveis pela execução de arquivos de mídia, os chamados *media players*, já se direcionaram uma resposta a esse problema. Por isso, a biblioteca de mídia, tornou-se uma função relativamente comum desse tipo de programa.

Figura 4 – Biblioteca de música digital



Fonte: autor

A ilustração acima apresenta a costumeira área da interface de um *media*

player, chamada de biblioteca de mídia. Ela exibe um campo de busca integrado a uma planilha cujas colunas remetem aos dados antes mencionados para classificar as músicas em pastas e subpastas. Esse tipo de programa traz, como qualquer evolução tecnológica, novas funcionalidades e possibilidades, as quais serão exploradas na seção seguinte.

3 REVISÃO DE DOCUMENTOS

A pesquisa documental foi realizada em caráter complementar aos pontos onde as obras teóricas se mostraram escassas. As fontes primárias consultadas consistiram nas documentações produzidas pelos próprios responsáveis por programas e bases de dados relacionados a coleções pessoais de mídia digital. Todavia, nesta seção, não se deixou as fontes secundárias completamente de lado. Elas foram empregadas sempre que disponíveis e pertinentes aos tópicos em questão.

3.1 FERRAMENTAS DE CATALOGAÇÃO DE MÍDIA DIGITAL

No cenário das **coleções institucionais** de mídia digital, segundo Suleman (2006), há uma demanda por ferramentas para gerenciar e tornar mais acessíveis, essas grandes e crescentes coleções de dados e metadados. Para esse autor, no entanto, tais ferramentas não estão prontamente disponíveis e, além disso, sistemas de bibliotecas digitais populares nem sempre se dimensionam adequadamente a coleções de diferentes tamanhos. Contudo, no cenário das coleções pessoais de mídia digital, em contraponto a afirmação de Suleman, há ferramentas prontamente disponíveis para gerenciar e tornar mais acessíveis as **coleções pessoais** de mídia digital.

De modo geral, essas ferramentas requerem do usuário que ele apenas indique as pastas a serem verificadas e elas, automaticamente, identificam e catalogam o conteúdo dos arquivos encontrados. Para realizar essa tarefa, esses *softwares* primeiro executam uma varredura nessas pastas e subpastas, elencando os arquivos que elas contêm. A seguir, há três maneiras distintas pelas quais a catalogação do conteúdo é iniciada:

a) a partir dos metadados embutidos nos arquivos;

b) a partir do nome do arquivo;

c) a partir do nome da pasta onde o arquivo está gravado.

No caso das informações catalográficas disponíveis estarem incompletas ou fora de padrão, esse tipo de programa pode corrigi-las importando os dados necessários de uma base informacional disponível na Internet (TAG, 2011). Dependendo do *software*, essas novas informações podem ser gravadas nos metadados do arquivo ou, apenas, registradas em uma base de dados local simplificada. A vantagem resultante de se empregar uma base de dados *off-line* se revela em uma maior velocidade na consulta às informações da coleção. Pois do contrário, cada busca demandaria uma nova varredura nas pastas e arquivos.

Outro fator comum às ferramentas de catalogação de coleções pessoais é a aderência, em maior ou menor grau, aos padrões para elementos de descrição definidos pelo International Standard for Bibliographic Description (ISBD):

a) título e indicação de responsabilidade;

b) edição;

c) publicação;

d) descrição física;

e) série;

f) notas;

g) número normalizado.

Conforme o procedimento de Spiteri (2009) – apresentado em seu estudo sobre os *sites* de catalogação social – tais elementos podem ser divididos em suas partes componentes da seguinte forma:

a) título;

b) autor;

c) edição;

d) lugar de publicação;

e) editora;

f) data de publicação;

g) extensão;

h) dimensão;

i) série;

j) nota;

k) ISBN.

A quantidade desses elementos empregada para catalogar os itens das coleções pessoais pode mudar de um *software* para outro. Da mesma forma, de um tipo de mídia para o outro é possível, também, que varie o nome atribuído a cada um desses elementos. Por exemplo, no caso de livros, uma “editora” é a responsável pela publicação, enquanto que no de uma música, esse nome muda para “gravadora” e, no de um filme, seria “estúdio”. O mesmo pode ser dito sobre o número normalizado, que pode ser um International Standard Book Number (ISBN), um International Standard Recording Code (ISRC) ou um International Standard Audiovisual Number (ISAN). A despeito dessas variações, a natureza informacional do elemento catalográfico ainda remete ao ISBD, o padrão internacional para descrições bibliográficas.

Entretanto, é invariável a finalidade a que se prestam essas ferramentas: facilitar a localização e a reprodução das mídias digitais. Em conformidade com o descrito no capítulo Metodologia, as próximas subseções apresentam os programas de catalogação eletrônica selecionados como objetos da pesquisa, através dos quais o aspecto da interoperabilidade foi analisado no capítulo 5.

A seguir, são apresentadas três ferramentas que realizam o processo automatizado de catalogação de mídias digitas.

3.1.1 Ferramenta para catálogos de músicas: Clementine

Uma das características mais interessante de projetos de *software* livre é a liberdade que é dada a qualquer interessado de criar o seu próprio projeto derivado. Um exemplo disso é o caso do Clementine. Quando o desenvolvimento do *music player* Amarok tomou um direcionamento completamente diferente – daquele que vinha apresentando até a versão 1.4 – David Sansome, John Maguire e Arnaud

Bienner preferiram criar seu próprio projeto, levando a diante a concepção anterior (JENKINS, 2011).

Em seu *website* o Clementine é descrito como tendo as seguintes características (JENKINS, 2011):

a) busca e reprodução de músicas da biblioteca nos formatos MP3, OGG, FLAC, MPC, M4A, AAC, WMA, MP4, SPX, WAV;

b) execução de rádios da Internet Spotify, Grooveshark, Last.fm, SomaFM, Magnatune, Jamendo, SKY.fm, Digitally Imported, e Icecast;

c) importação e exportação de listas de execução nos formatos M3U, XSPF, PLS, ASX e CUE;

d) visualizações com o projectM;

e) exibição de letras de música, biografia de artistas e fotos;

f) conversão de música em MP3, OGG VORBIS, SPEEX OGG, FLAC ou AAC;

g) obtenção de capas para os álbuns a partir do Last.fm e Amazon;

h) transferência das músicas para iPod, iPhone, ou qualquer *player* com armazenamento compatível com USB ou MTP.

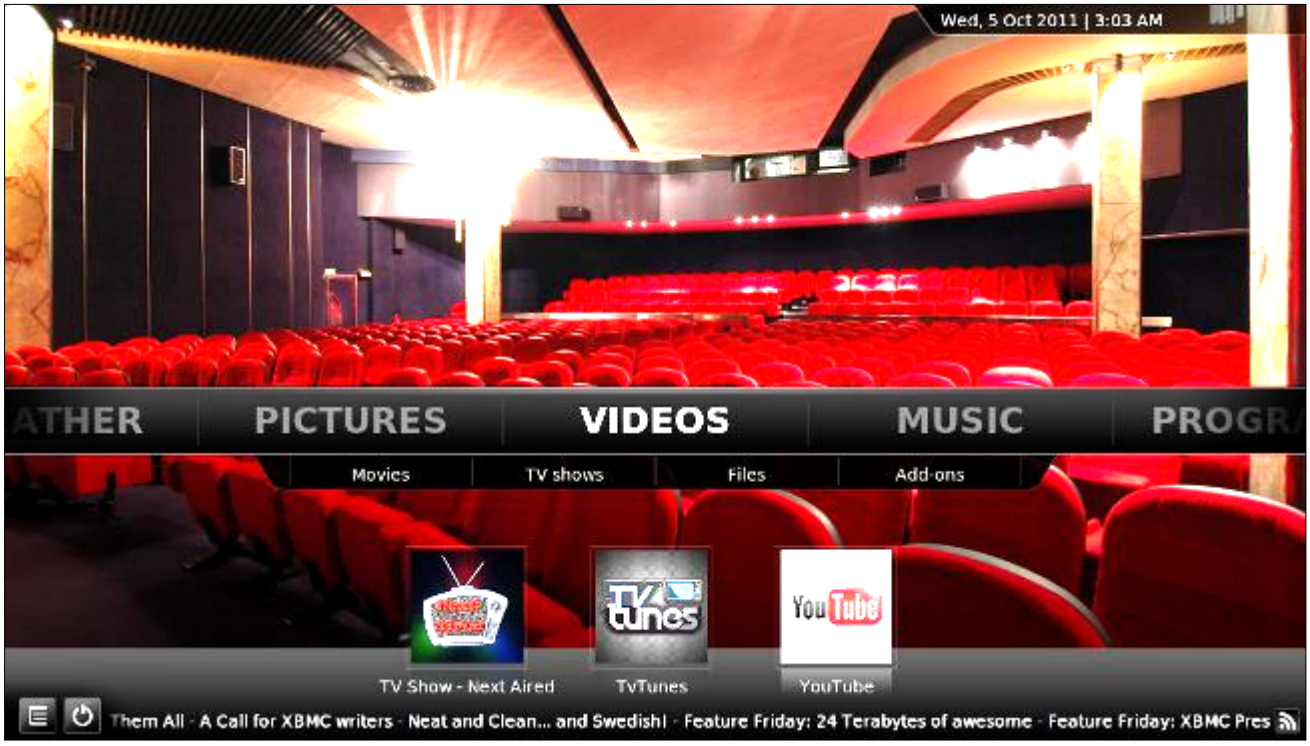
O Clementine administra coleções de arquivos em formatos de compressão de áudio com perda (MP3) e sem perda dados (FLAC). Esse é um fator importante para audiófilos que possuem aparelhagem para reprodução de som com alta fidelidade (*hi-fi*). No caso do dono da coleção preferir ouvir suas músicas em aparelhos digitais portáteis, ele gerencia a transferência do computador para o dispositivo (alínea “h”).

Contudo, esse *software* não se restringe a apenas organizar e reproduzir os arquivos de música do computador onde está instalado. Ele é capaz de se conectar pela Internet a transmissões de rádios digitais (alínea “b”), e oferecer o mesmo tipo de funcionalidades disponíveis às músicas que pertencem à coleção. Ou seja, o Clementine consegue exibir informações como letras, biografia, fotos dos artistas (alínea “e”) e visualizações animadas no ritmo da música (alínea “d”) que a rádio estiver tocando.

3.1.2 Ferramenta para catálogos de vídeos: XBMC

Entre os programas que cumprem com a finalidade de facilitar a localização e execução de vídeos digitais estão os chamados *media centers*. A principal diferença entre eles e os *media players* está na interface gráfica, semelhante a dos menus dos filmes em DVD. Esse tipo de interface é, normalmente, planejada para poder ser usada a uma distância de 3m da tela e ser operada com um controle-remoto, não com um teclado e um mouse (10-FOOT, 2012).

Figura 5 – Tela inicial de um *media center*



Fonte: xbmc.org

A ilustração acima corresponde a tela inicial do *media center* XBMC. Conforme explica a sua documentação (10-FOOT, 2012), através dessa interface o usuário pode facilmente “navegar” por sua coleção e acessar seus vídeos, fotos e música usando apenas os botões de um controle-remoto. O XBMC é um projeto desenvolvido por voluntários e gerenciado por uma organização sem fins lucrativos chamada XBMC Foundation.

Esse *software* foi inicialmente criado para ser instalado no *hardware* do

videogame Xbox, fabricado pela Microsoft, e aí está o significado original do nome XBMC (Xbox Media Center). Contudo, suas versões mais recentes deixaram de suportar essa instalação em nome de uma melhor integração com os sistemas operacionais Linux, OS X e Windows (XBMC, 2012). O resultado disso é que, mesmo computadores, equipados com processadores e placas de aceleração gráficas menos potentes, são capazes de executar essa ferramenta – variando, proporcionalmente, então, a resolução do vídeo a ser reproduzido. Quanto mais potente for o equipamento, maior poderá ser a definição do vídeo.

Ainda seguindo as informações presentes em sua documentação, sua versão atual é capaz de reproduzir todos os formatos de vídeo e áudio mais populares (XBMC, 2012). Essa afirmação pode ser verificada analisando-se parte da sua lista de formatos suportados (FEATURES, 2012):

a) formatos de vídeo: MPEG-1, MPEG-2, H.263, MPEG-4 SP e ASP (Xvid/DivX), MPEG-4 AVC (H.264), HuffYUV, Indeo, MJPEG, RealVideo, RMVB, Sorenson, WMV, Cinepak;

b) formatos de áudio: MIDI, AIFF, WAV/WAVE, AIFF, MP2, MP3, AAC, AACplus (AAC+), Vorbis (Ogg), AC3, DTS, ALAC, AMR, FLAC, Monkey's Audio (APE), RealAudio, SHN, WavPack, MPC/Musepack/Mpeg+, Shorten, Speex, WMA, IT, S3M, MOD (Amiga Module), XM, NSF (NES Sound Format), SPC (SNES), GYM (Genesis), SID (Commodore 64), Adlib, YM (Atari ST), ADPCM (Nintendo GameCube), e CDDA;

c) formatos de contêiner: AVI, MPEG, WMV, ASF, FLV, MKV/MKA (Matroska), QuickTime, MP4, M4A, AAC, NUT, Ogg, OGM, RealMedia, 3gp, VIVO, PVA, NUV, NSV, NSA, FLI, FLC, DVR-MS e WTV;

d) formatos de legendas: AQTitle, ASS/SSA, CC, JACOsub, MicroDVD, MPsub, OGM, PJS, RT, SMI, SRT, SUB, VOBsub, VPlayer;

e) protocolos de rede: AirPlay/AirTunes, UPnP, SMB, AFP, DAAP, Zeroconf, NFS, HTTP, HTTPS, FTP, RTSP, MMS, Podcasting, TCP, UDP, SFTP, RTP e RTMP, DHCP, NTP, JSON-RPC, D-Bus.

Os desenvolvedores desse *software*, evidentemente, preocuparam-se em permitir que seus usuários tivessem a possibilidade de reproduzir uma grande variedade de padrões de áudio (alínea “a”) e vídeo (alínea “b”). Essa variedade corresponde ao número de possíveis combinações, de um tipo de vídeo com um tipo de áudio, o qual é limitado apenas pelo formato escolhido para servir de contêiner

(alínea “c”). Nem todos os formatos de contêineres – espécie de “embalagem” para as faixas de áudio, vídeo e legenda – são compatíveis com todos os formatos para imagens, sons e texto. Por exemplo, o arquivo de um filme gravado no formato de contêiner AVI, pode ter como formato de vídeo o Xvid e, como formato de áudio, o MP3, mas não o formato Ogg Vorbis (MATROSKA, 2011). E, se o usuário, desejar assistir a esse filme acompanhado de legendas, ele pode usar um arquivo de texto independente do arquivo do filme, desde que esse esteja em qualquer dos formatos citados na alínea “d”.

Como se constata pela alínea correspondente aos protocolos de rede, o XBMC é capaz de reproduzir conteúdo multimídia através de praticamente qualquer protocolo disponível (XBMC, 2012). Isto permite que ele gerencie, em sua base de dados, arquivos distribuídos em diferentes aparelhos da casa – criando uma biblioteca de mídia digital doméstica.

3.1.3 Ferramenta para catálogos de livros: Calibre

O Calibre, segundo seu manual (GOYAL, 2012), é um *software* gratuito e de código aberto para gerenciamento de bibliotecas de livros eletrônicos (*e-books*). O início de seu desenvolvimento remonta a 2006, quando Kovid Goyal percebeu que seu leitor de livros eletrônicos (*e-book reader*), recém adquirido, não era compatível com o sistema operacional Linux que utilizava. Então, decidiu criar um sistema que permitisse a transferência de arquivos entre seu computador e o novo aparelho. A esse sistema foi agregado um conversor de formatos de *e-books* e, mais tarde, conforme crescia a sua coleção e mais complicado se tornava organizá-la, foi adicionada uma interface gráfica para facilitar tal tarefa (GOYAL, 2009). Hoje, suas principais funcionalidades são descritas da seguinte maneira:

a) administração de biblioteca digital;

b) conversão de *e-books;*

c) sincronização da coleção com dispositivos de leitura de *e-books*;

d) obtenção de notícias da *web* e conversão delas em *e-book*;

e) visualizador de *e-books*;

f) servidor para acesso *on-line* a coleção pessoal.

Dentre as fontes brasileiras de notícias da Internet, que podem servir para conversão em e-books, estão listadas: ClicRBS, O Globo, Folha de São Paulo e O Estado de São Paulo. Essa conversão é possível porque o HTML está entre os formatos que o Calibre suporta para entrada de dados. A lista completa presente em sua documentação é a seguinte (GOYAL, 2012):

a) formatos de entrada: CBZ, CBR, CBC, CHM, DJVU, EPUB, FB2, HTML, HTMLZ, LIT, LRF, MOBI, ODT, PDF, PRC, PDB, PML, RB, RTF, SNB, TCR, TXT, TXTZ;

b) formatos de saída: AZW3, EPUB, FB2, OEB, LIT, LRF, MOBI, HTMLZ, PDB, PML, RB, PDF, RTF, SNB, TCR, TXT, TXTZ.

Além do formato típico de páginas da Web, são aceitos como entrada, desde formatos de arquivos complexos, como os de suítes de escritório (ODT, RTF e PDF), passando por formatos específicos para histórias em quadrinhos (CBZ, CBR e CBC), até simples arquivos de texto puro (TXT). A partir deles, podem ser gerados arquivos para leitura nos mais populares *e-book readers* da atualidade (AZW3 para o Kindle e EPUB para o IPAD) ou simplesmente transformá-los em formatos com os quais o Calibre trabalha melhor (LIT e MOBI), para visualização apenas na tela do computador.

Associado a sua capacidade de organização de uma biblioteca de livros digitais, o Calibre, também, pode transformar o computador onde for instalado em uma espécie de servidor Web. Isso permite que o dono da coleção possa acessa-lá remotamente através da Internet – usando um navegador – além de compartilhá-la com quem desejar.

3.2 BASES UTILIZADAS PELAS FERRAMENTAS DE CATALOGAÇÃO

Nas subseções seguintes serão apresentadas bases de dados *on-line* que fornecem informações referenciais sobre músicas, filmes e livros. A seleção destas bases fundamentou-se no fato de permitirem que seus usuários finais editem e incluam novos dados, alimentando a base. Além de serem ilustrativas do tipo de

fonte de informação que é utilizado pelas ferramentas de catalogação mencionadas na seção anterior.

3.2.1 Bases sobre músicas: Gracenote e MusicBrainz

A Gracenote é a companhia antigamente conhecida como Compact Disc Data Base (CDDB). Sua proposta de negócio é manter e licenciar uma base de dados disponível através da Internet, contendo informações sobre conteúdo de áudio, seja em CDs ou gravações em vinil. Hoje ela fornece programas e metadados para empresas que disponibilizam a seus consumidores a possibilidade de busca e gestão de mídia digital. “Com mais de 250 milhões de consumidores usando mais de 2 mil aplicativos cadastrados na Gracenote, a base de dados global de mídia processa dezenas de milhares de buscas diárias e bilhões por mês” (GRACENOTE, 2011).

Para agregar essa quantidade de informações é empregada uma combinação de quatro métodos diferentes:

a) alimentação direta da base pelos mais de 3 mil detentores de conteúdo (gravadoras, editoras e estúdios de cinema);

b) uso de especialistas em música e vídeo para classificação e categorização por gênero e tipo de mídia;

c) audição automatizada por *software* para classificar músicas e aplicar descritores;

d) incorporação das contribuições da comunidade de usuários dos programas e serviços cadastrados na Gracenote.

Oferecendo o mesmo tipo de serviço, mas em versão gratuita, destaca-se o projeto MusicBrainz. Ele foi fundado em resposta às restrições de acesso a base da CDDB quando essa se tornou a Gracenote. Segundo o próprio MusicBrainz (2011), seus dois principais objetivos são:

a) tornar-se a melhor fonte de informação musical, liberando os dados sob licenças abertas e permitindo que qualquer pessoa contribua;

b) tornar-se uma língua franca universal para música, fornecendo uma forma confiável e inequívoca de identificação musical, permitindo que pessoas e

máquinas tenham conversas significativas sobre música.

Sua base de dados é mantida exclusivamente por voluntários que seguem os padrões propostos pelo projeto. Usuários finais podem utilizar programas que se comunicam com essa base e, assim, importar metadados para arquivos de áudio digital em diversos formatos.

3.2.2 Bases sobre filmes e seriados: IMDB e TheTVDB

A Internet Movie Database (IMDB), atualmente propriedade da Livraria Amazon, é uma base de dados de informação relativa a filmes, seriados de televisão, videogames e demais dados relativos as suas produções. Conforme seu *site,* desde 1990, pessoas do mundo todo têm adicionado, refinado e corrigido os dados da base. Cada informação submetida ao IMDb é verificada antes ser disponibilizada *on-line* (INTERNET MOVIE DATABASE, 2011).

A contrapartida livre do IMDB é o The Television Database (TheTVDB). De acordo com seu website sua meta é tornar-se a mais completa e precisa fonte de informação sobre séries de televisão de qualquer língua ou país. Além disso, também fornece um repositório de imagens de temporadas e episódios de seriados que podem ser integrados a *software*s de *Home Theaters* (THE TELEVISION DATABASE, 2011).

Essa é uma base de dados que pode ser alterada por qualquer pessoa, sem verificação prévia. Todas as imagens e informações vêm da contribuição dos usuários. O The TVDB fornece ferramentas de integração de sua base de dados permitindo agregar suas informações em outros programas.

3.2.3 Bases sobre livros: LibraryThing e Anobii

A LibraryThing é uma aplicação baseada em Web voltada para catalogação social, através da qual seus usuários podem armazenar e compartilhar os

metadados de seus livros. Ela é usada tanto por leitores quanto por autores, editoras e livrarias. Sua principal característica é colocar em contato pessoas com os mesmos livros e fornecer sugestões de leitura (LIBRARYTHING, 2011).

Outro *site* com finalidade semelhante é o Anobii. Ele se caracteriza como uma rede social voltada para leitores. Seu serviço permite que indivíduos cataloguem e deem notas aos seus livros, disponibilizando espaço para publicarem resenhas e discutirem seus textos. Segundo seu *website* a sua missão é juntar os amantes dos livros e encorajar a leitura (ANOBII, 2011).

4 METODOLOGIA

Nesta seção são apresentadas e explicadas as decisões referentes à metodologia aplicada na realização dos objetivos do presente estudo.

4.1 NATUREZA E ABORDAGEM

A intenção que permeia o desenvolvimento deste estudo é a de produzir conhecimento, voltado para aplicações práticas em projetos de consultoria que envolvam coleções pessoais de mídia digital. Isso caracteriza a presente pesquisa como sendo de natureza aplicada.

A abordagem ao problema escolhido foi a qualitativa. Essa estratégia de estudo é a que melhor atende a necessidade de descrição da complexa interação dos elementos envolvidos no funcionamento dos diferentes modelos de interoperabilidade.

4.2 MÉTODO DA PESQUISA

O procedimento metodológico escolhido para a realização deste projeto foi a combinação da pesquisa documental com a pesquisa bibliográfica. Para Marconi e Lakatos (2006, p. 62) a pesquisa documental “[. . .] está restrita a documentos, escritos ou não, constituindo o que se denomina de fontes primárias”, enquanto que na bibliográfica recorre-se às fontes secundárias através da consulta a “[. . .]um apanhado geral sobre os principais trabalhos já realizados [. . .] sobre o tema” (LAKATOS; MARCONI, 1991, p. 158).

Santos e Candeloro (2006, p.70), afirmam que a “[. . .] pesquisa bibliográfica consiste na busca de elementos para sua investigação em materiais impressos ou editados eletronicamente”, não se pode negar que o mesmo se aplica a pesquisa

documental, a despeito da classificação da fonte sobre a qual se dá. Assim adotou-se a mesma metodologia, sugerida por Macedo (1996), para lidar com os dois tipos de fontes:

a) identificar, localizar e obter documentos pertinentes ao estudo, criando uma bibliografia básica;

b) elaborar um esquema provisório dos temas e subtemas, arrolando os principais descritores;

c) enriquecer a bibliografia básica, realizando uma nova busca com descritores não usados inicialmente.

A pesquisa bibliográfica se deu inicialmente através da consulta a material da literatura especializada na área. Essas fontes secundárias foram obtidas através de busca nas seguintes bases de dados:

a) Sistema de Bibliotecas da Universidade Federal do Rio Grande do Sul;

b) Portal de Periódicos da CAPES,

c) Scientific Electronic Library Online (SciELO),

d) Google Acadêmico;

e) Lecture Notes in Computer Science (LNCS);

f) Directory of Open Access Journals (DOAJ);

g) ProQuest Dissertations & Theses A&I.

O material encontrado na pesquisa bibliográfica cumpriu com o primeiro objetivo específico deste estudo: localizar temas relacionados a esta pesquisa na literatura especializada nacional e estrangeira. A seguir, as informações dessas fontes secundárias foram complementas pela pesquisa documental, que se baseou nos dados disponibilizados pelos criadores ou responsáveis pelas ferramentas de catalogação eletrônica e bases de dados informacionais.

4.3 OBJETO DA PESQUISA

O objeto da pesquisa circunscreve-se aos modelos de interoperabilidade das ferramentas de catalogação eletrônica de coleções pessoais de mídia digital. Isto engloba os três vetores do processo de intercâmbio de informação. O primeiro, a base de dados *on-line*, o segundo o modelo de transferência e, por fim, a ferramenta que cataloga a coleção.

Para melhor compreender como esses modelos são aplicados na prática, foram selecionados três programas que apresentam o aspecto da interoperabilidade. Cada um deles destinado à catalogação de um tipo diferente de mídia digital:

a) Calibre, para livros;

b) Clementine, para músicas;

c) XBMC, para vídeos.

Esses *softwares* de catalogação eletrônica foram escolhidos pela sua conformidade aos seguintes critérios:

a) empregam um ou mais modelos de interoperabilidade;

b) funcionam nativamente nos sistemas operacionais Windows, OS X e Linux.

A explicação para esse segundo quesito está ligada ao fato de que um programa que funcione nesses três sistemas operacionais tem mais possibilidade de ser aproveitado em um leque maior de projetos, se comparado a um programa compatível apenas com um único sistema. Portanto, esses são *softwares* potencialmente mais relevantes a um bibliotecário consultor.

4.4 INSTRUMENTO DE COLETA DOS DADOS

Para fins de padronização da coleta de dados, foi empregada a técnica de formulários. Ela permitiu avaliar e comparar descritivamente o funcionamento dos modelos de interoperabilidade implementado pelos diferentes programas.

O formulário constituiu-se de uma lista formalizada de tópicos a serem averiguados sobre os *softwares* selecionados. O relacionamento, de cada tópico dessa lista, com os objetivos deste trabalho, é ilustrado a seguir:

Quadro 1 – Relação dos itens do formulário com os objetivos

OBJETIVO ESPECÍFICO	**FORMULÁRIO DE COLETA DE DADOS SOBRE MEIOS DE INTEROPERABILIDADE EM PROGRAMAS PARA CATALOGAÇÃO DE COLEÇÕES PESSOAIS DE MÍDIA DIGITAL**
	IDENTIFICAÇÃO DO PROGRAMA
Identificar programas de catalogação eletrônica orientados a música, vídeos e livros digitais	1 Nome:
	2 Versão:
	3 Data de lançamento inicial:
	4 Tipo de licença:
	5 Tipos de mídia que cataloga:
	MEIOS DE INTEROPERABILIDADE
Identificar e caracterizar as fontes utilizadas pelos processos automatizados de catalogação em coleções pessoais.	6 Fontes de informação utilizadas:
Descrever modelos de interoperabilidade utilizados nos processos automatizados de catalogação em coleções pessoais.	7 Nível de funcionalidades e custos do modelo de interoperabilidade:
Avaliar os modelos de representação dos metadados de acordo com o padrão internacional para descrição bibliográfica (ISBD).	8 Adequação do formado dos metadados ao ISBD:
Descrever o processo automatizado de catalogação em coleções pessoais de música, vídeos e livros.	9 Método de identificação do item da coleção:
	10 Local onde são salvos os novos metadados:

Fonte: autor.

5 APRESENTAÇÃO E ANÁLISE DAS FERRAMENTAS

Nas subseções seguintes serão apresentadas e analisadas as informações relativas aos programas de catalogação eletrônica .

5.1 IDENTIFICAÇÃO DOS *SOFTWARES*

Para facilitar a comparação, os dados coletados para identificação dos *software*s foram agrupados no seguinte quadro:

Quadro 2 – Identificação dos *softwares*

NOME:	**CLEMENTINE**	**XBMC**	**CALIBRE**
Versão:	1.0.1	11.0	0.8.55
Ano do lançamento inicial:	2010	2003	2006
Tipo de licença:	GPL	GPL	GPL
Tipo de mídia que cataloga:	música	vídeo, música e imagem	livro

Fonte: autor.

O primeiro ponto a ser comentado refere-se ao estado atual das versões desses programas. Calibre é o único deles cujo desenvolvimento ainda não atingiu o marco da versão “1”. Normalmente, projetos de *software* livre costumam reservar esse número de versão para aquele lançamento que inclui todas as funcionalidades que se pretendia desenvolver, estando, então, o *software* “completo”. Contudo, isso, de forma alguma, significa que ele não estaria estável e pronto para uso – significa, apenas, que ainda há funções a serem implementadas.

O ano do lançamento inicial mostra a quanto tempo o *software* está no

“mercado”. Isso sinaliza o grau de maturidade e confiabilidade do projeto como um todo. Neste caso, pode se inferir que é menos provável que o desenvolvimento do XBMC seja interrompido, do que o do Clementine – sendo este muito mais recente do que aquele.

Quanto ao tipo de licença, os três *software*s usam o mesmo tipo de licença livre. O que, para o usuário, em termos práticos, significa que os três podem ser usados gratuitamente, porém sem qualquer garantia que responsabilize os desenvolvedores por algum defeito ou mal-funcionamento.

Quanto ao tipo de mídia que cada um cataloga, apenas o XBMC se dedica a mais de um tipo – podendo catalogar, além de vídeos, também, imagens e músicas. Isso amplia a variedade de projetos de organização de acervos onde ele pode ser aproveitado. Todavia, sua avaliação neste trabalho concentrou-se apenas no seu tipo principal de mídia.

5.2 MEIOS DE INTEROPERABILIDADE

As subseções seguintes apresentam a análise dos dados relativos aos métodos, modelos e padrões que servem de meios para a realização da interoperabilidade no processo automatizado de catalogação – realizado pelas ferramentas identificadas na seção anterior.

5.2.1 Fontes de informação utilizadas para catalogação

As fontes de informação que cada uma dessas ferramentas pode consultar, durante o processo automático que gera o catálogo para a coleção pessoal, são as apresentadas no quadro seguinte:

Quadro 3 – Fontes de informação para catalogação

CLEMENTINE	XBMC	CALIBRE
7digital	Amazon	Adlibris.com
Amazon	AniDB.net	Amazon
Aol Music	AsianDB	Anobii
AZLyrics.com	AtMovies	Antikvarium.hu
Dark Lyrics	BestAnime	Baen Ebook
DirectLyrics	Beyazperde	Barnes & Noble
Discogs	Cinefacts	Beam eBooks
eLyrics.net	Cinemagia	BiblioEteca
Facebook	Cinemarx	Bol.com
iTunes	Ciné-Passion	Bol.de
Last.fm	Daum Movie	Catawiki
Letras.mus.br	Disney Video NL	Databaze Knih
LoudSon.gs	Douban	Dizzie.nl
Lyrics Mania	FDB.pl	Douban
Lyrics Plugin	FilmAffinity	Fantastic Fiction
Lyrics.com	FilmBasen	FictionDB
LyricsBay	FilmDelta	Goodreads
LyricsMode.com	FilmStarts	Google Books
LyricsMode.com	Filmweb	ISBNDB.COM
LyricsReg.com	GetLib	LibraryThing
LyricsTime.com	IMDB	Libri.hu
LyricWiki	Kino.de	Lubimy Czytać
Lyriki	KinoPoisk	Moly.hu
MetroLyrics	Kinobox.cz	Overdrive
MOG	MTime	Ozon.ru
Mp3 Lyrics	MovieMaze	Skoob
Mtv Music	MovieMeter	
MusicBrainz	MoviePlayer.it	
MySpace	MyMovies.dk	
Rdio	MyMovies.it	
Seek Lyrics	Naver Movie	
Song Lyrics	OFDB	
Tekstowo.pl	PTGate	
Teksty.org	Port.hu	
Vagalume	Rotten Tomatoes	
Wikipedia	SerialZone.cz	
	Sratim	
	The MovieDB	
	The TVDB	
	VideoBuster	
	WorldArt	
	Yahoo!Taiwan Movie	

Fonte: autor.

As fontes listadas acima são de natureza bastante variada. A lista não inclui

somente bases de dados, no sentido mais convencional (como IMDB e The TVDM), mas também redes de relacionamento (como Facebook e MySpace), sites de comércio eletrônico (como Amazon e Barnes & Noble) e sites de catalogação social (como LibraryThing e Douban).

Considerando os dois tipos principais de bases de dados – as de fontes e as de referência – citados por Rowley e Hartley (2008), não se pode dizer, em sentido estrito, que qualquer desses *sites* se enquadre nessa tradicional categorização. Contudo, ao ser relativizada, essa classificação pode ser empregada para analisar o modo como as ferramentas de catalogação se relacionam com cada uma de suas fontes de informação. Por exemplo: o Clementine utiliza o *site* brasileiro Vagalume como se ele fosse uma base de dados de fontes, de onde ele recupera um documento que não precisa ser complemento, o texto integral da letra de uma música. Em contrapartida, o MusicBrainz poderia ser classificado como uma fonte de referência para o Clementine, já que dele são recuperados apenas metadados, os quais, pela própria definição, apontam para outro documento.

Desse modo, seria possível categorizar cada uma de suas fontes de informação segundo o tipo de uso que cada ferramenta faz dela. Contudo, a documentação disponibilizada pelas ferramentas raramente atinge esse nível de detalhamento. O que demandaria testes de *software* que não se enquadram no escopo do presente trabalho.

5.2.2 Nível de funcionalidades e custos do modelo de interoperabilidade

Quando se desenvolve uma ferramenta de catalogação para coleções institucionais, o custo e as funcionalidades constituem parâmetros fundamentais na escolha de uma abordagem para a questão da interoperabilidade. Da mesma forma, no que se refere à seleção de uma ferramenta para catalogação de coleções pessoais, esses fatores são cruciais na escolha de um *software* que melhor atenda às necessidades do usuário do acervo digital.

Considerando-se a lista de tipos de abordagens à interoperabilidade,

apresentada por Arms (2000) e aqui comentada na seção 2.2, é possível fazer a seguinte classificação:

a) raspagem de dados: poucas funcionalidades e custo baixo;

b) busca federada de informações: muitas funcionalidades e custo alto;

c) colheita de metadados: importantes funcionalidades e custo moderado.

A questão dos custos, em um ambiente de *software* livre, se traduz principalmente em tempo. O quanto demora a implementação da funcionalidade ou para repará-la em caso de alterações inesperadas nas fontes dos dados.

Cada uma das ferramentas analisada no presente estudo faz uso de pelo menos um desses modelos para obtenção de metadados descritivos ou informações complementares. O quadro seguinte relaciona a função que é atribuída a cada modelo em cada uma das três ferramentas:

Quadro 4 – Modelo de interoperabilidade por ferramenta

	CLEMENTINE	**XBMC**	**CALIBRE**
Usa busca federada para:	Letra de músicas. Biografia de artista.	-	Capas.
Usa colheita para:	-	-	Notícias.
Usa raspagem:	-	Metadados. Pôsteres. Notas.	Metadados.

Fonte: autor.

No Clementine, a busca federada é empregada para recuperar, tanto a biografia dos artistas, quanto a letra das músicas que estiverem em execução. Isso permite ao usuário escolher visualizar o resultado, já padronizado, da fonte que preferir. O mesmo sistema é adotado pelo Calibre para obter as capas para os livros digitais.

O Calibre é o único dos três *software*s a empregar os três modelos de interoperabilidade. Além da busca federada, ele utiliza a colheita para acessar *websites* de jornais – em datas e horários pré-agendados – recuperando notícias

atualizadas para convertê-las em *e-books*; e, a raspagem para completar os metadados dos itens da coleção.

No caso do XBMC, a raspagem é empregada, não só para completar metadados descritivos (como título e diretor), mas também para obter informações complementares (como pôsteres de filmes e de seriados, além de notas atribuídas a eles pelo voto de usuários de sites como o IMDB e The TVDB). O único dos três *softwares* a não usar esse modelo para busca e obtenção de metadados é o Clementine, pois ele adota a API do MusicBrainz para essa função.

5.2.3 Adequação do formato dos metadados ao ISBD

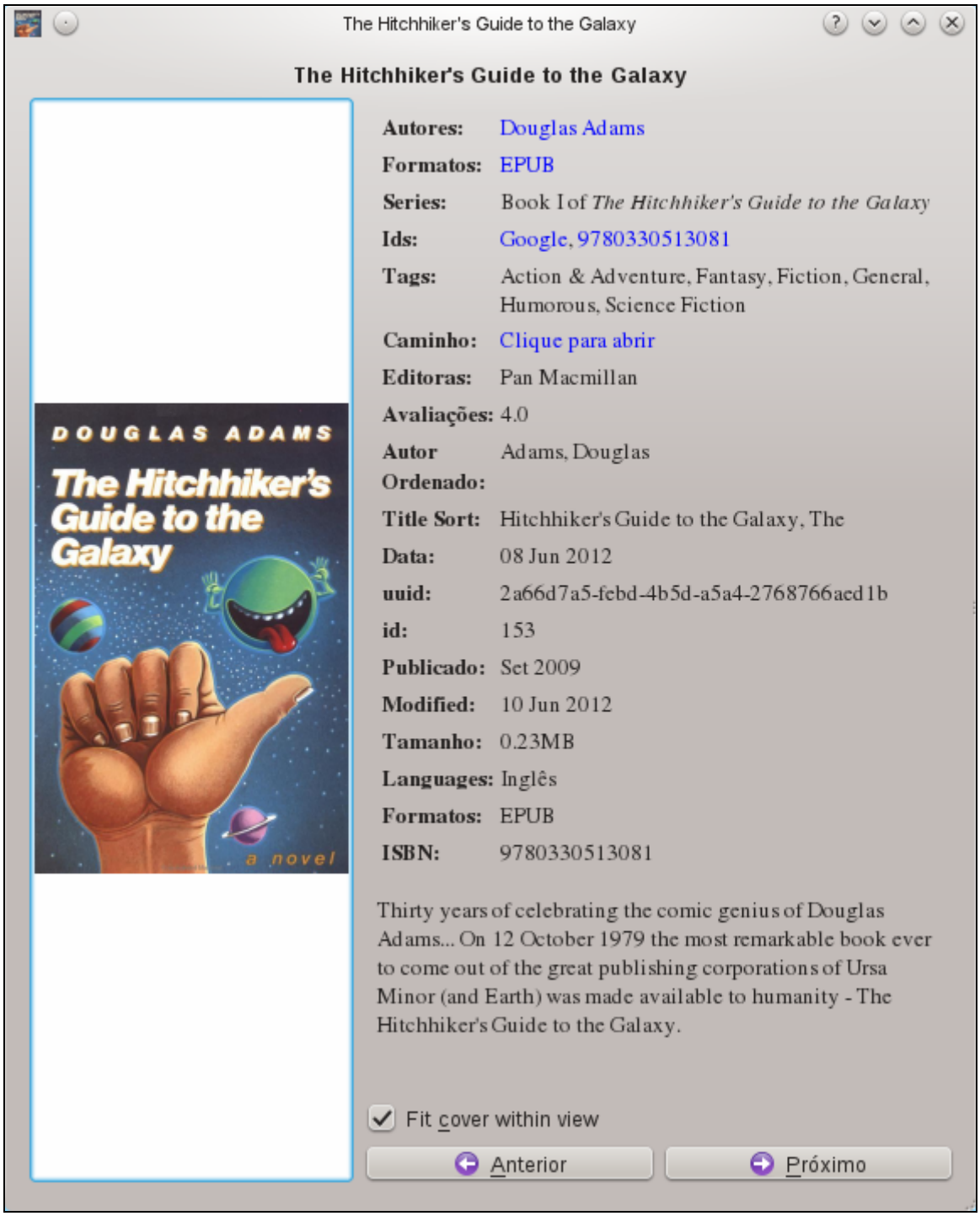
O padrão internacional para descrição bibliográfica (ISBD) lista os elementos pertinentes a um bom registro catalográfico. Por isso é importante verificar se, durante o processo automatizado de catalogação, as ferramentas contemplam todos esses elementos. E, em caso negativo, é preciso identificar quais deles não estão presentes.

Figura 6 – Descrição de música no Clementine

Fonte: autor

Como visto nas duas telas que compõem a ilustração anterior, as informações catalográficas do Clementine não contemplam as áreas para: edição, lugar de publicação, gravadora, série e número padronizado.

Figura 7 – Descrição de livro no Calibre



Fonte: autor

No caso do Calibre também há campos faltando, mas neste caso são os que informam a edição, o lugar de publicação, a extensão.

Figura 8 – Descrição de filme no XBMC

Fonte: autor

Na descrição apresentada pelo XBMC, faltaram as informações relativas às áreas de edição, lugar de publicação e número padronizado.

A lista de metadados descritivos, que não foram incluídos no sistema de catalogação de cada uma das três ferramentas, ilustra o quadro seguinte:

Quadro 5 – Elementos do ISBD que faltam em cada ferramenta

	CLEMENTINE	**XBMC**	**CALIBRE**
Elementos do ISBD que estão ausentes	edição lugar de publicação gravadora série número padronizado	edição lugar de publicação número padronizado	edição o lugar de publicação extensão

Fonte: autor

Apesar da existência de elementos do ISBD não contemplados na descrição bibliográfica elaborada automaticamente por todas as três ferramentas, nenhuma das ausências compromete a utilidade do *software* no cenário das coleções pessoais.

5.2.4 Método de identificação dos itens da coleção

Para realizar o processo automatizado de catalogação, as ferramentas precisam selecionar uma ou mais informações relativas ao item para utilizar como termo de busca nas bases de dados *on-line* – e, assim, obter os metadados descritivos. No quadro seguinte, está ilustrado o método adotado por cada uma das ferramentas aqui analisadas:

Quadro 6 – Método de identificação dos itens da coleção

	CLEMENTINE	**XBMC**	**CALIBRE**
Nome da pasta	X	X	X
Nome do arquivo		X	
Metadados no arquivo	X		X

Fonte: autor.

Como visto, o Clementine e o Calibre adotam o mesmo método, através do qual combinam as informações extraídas dos metadados do arquivo e do nome da pasta onde ele se encontra – sem considerar o nome do próprio arquivo. Os dados obtidos através da busca nas fontes são usados, então, para catalogar os arquivos como itens individuais ou como parte de um mesmo álbum ou de uma única série de livros.

Já o XBMC ignora os metadados descritivos presentes nos arquivos, empregando os nomes do arquivo e da pasta para obter novos metadados descritivos a partir das fontes de informação da Internet.

5.2.5 Local onde são salvos os novos metadados

Os metadados obtidos através da interoperabilidade com fontes de informações da Internet são salvos na coleção digital do usuário. O local, onde cada ferramenta salva esses dados, ilustra o quadro a seguir:

Quadro 7 – Local de salvamento dos novos metadados

	CLEMENTINE	**XBMC**	**CALIBRE**
Base de dados local	X	X	X
Base de dados on-line			
No arquivo do próprio documento	X		

Fonte: autor.

Para acelerar a busca de itens na própria coleção, era esperado que cada ferramenta salvasse em uma base de dados própria todos os metadados da coleção. Porém, somente o Clementine salva os novos metadados também no item da coleção, atualizando diretamente o arquivo. E nenhuma das ferramentas analisadas permite o caminho contrário. Caso os dados na fonte de informações *on-line* estiverem incompletos, o usuário das ferramentas não poderá complementá-los através delas.

6 CONSIDERAÇÕES FINAIS

Os bibliotecários sempre precisaram se adaptar aos novos suportes da informação. Eles fazem isso desde a época dos papiros e dos pergaminhos. Contudo, mesmo que o suporte agora não seja mais físico, as mídias digitais não se destacam por apresentarem grandes desafios, mas sim novas facilidades para o organizador dos acervos. O processo automatizado de catalogação consegue, em poucos minutos, transformar um enorme acervo – no qual o próprio dono tem dificuldade de encontrar qualquer coisa – em uma biblioteca de mídia bem organizada.

A vantagem desse método, para o catalogador, está no tempo ganho e na facilidade da execução automática, que depois de iniciada, se torna não interativa. Obviamente, essa última característica só é válida para os usuários que não precisam de nenhum dos elementos do ISBD que a ferramenta não contemplar. Pois, nesse caso, ainda restará a tarefa manual de incluir esses elementos nas áreas de notas ou comentários que a ferramenta reserva para cada item que ela cataloga.

De qualquer forma, em linhas gerais, os componentes do processo automatizado são três: a coleção digital, a fonte de informações *on-line* e os meios de interoperabilidade, que transferem os dados de uma para a outra. Dentre esses meios, foi nos modelos de interoperabilidade que o presente estudo se concentrou.

Tais modelos, empregados na obtenção de informações a partir de fontes *on-line*, foram classificados em: busca federada, colheita e raspagem. Para os dois primeiros modelos existem protocolos e padrões institucionais bem estabelecidos. Todavia, no âmbito não tradicional das coleções pessoais, mesmo esses modelos são implementados através de soluções de baixo custo não padronizadas, transferindo dados através da interface homem-máquina – principal característica da raspagem de dados.

A consequência disso é que, como não há um acordo prévio com os fornecedores de informação, a interface da fonte pode ser alterada sem aviso. Ou seja, no cenário das coleções pessoais, não há qualquer garantia de disponibilidade

para a interoperabilidade com as fontes de informação *on-line* – em um momento ela pode estar funcionando e, no outro, não estar mais. Nessa situação, cabe ao usuário escolher entre:

a) aguardar até que os desenvolvedores da ferramenta lancem uma correção para essa falha na comunicação com a fonte, ou;

b) usar outra fonte que forneça o mesmo tipo de informação.

Essa segunda opção só é possível porque as ferramentas dispõem de várias fontes que contêm o mesmo tipo de informação. Essa redundância de fontes atende justamente a necessidade de manter funcionando o processo de catalogação, mesmo sem uma ou outra fonte. Contudo, ainda que o tipo de informação seja o mesmo, isso não garante que a qualidade dela seja a mesma. Por exemplo: no caso de uma modificação na interface da Amazon, o processo de catalogação do Calibre ainda poderia continuar funcionando usando apenas as informações recuperadas do Google Books. Cabendo ao usuário decidir se a qualidade dos metadados descritivos dessa fonte lhe atende do mesmo modo como os recuperados daquela. Se a resposta for um não, então, talvez o usuário precise aguardar por uma atualização do Calibre.

Por outro lado, as fontes de informação empregadas nesse processo, nem sempre se caracterizam como bases de dados propriamente ditas. Portanto, a classificação tradicional de bases de dados, como sendo de fontes ou de referência, não se aplica perfeitamente, precisando ser relativizada. Assim, uma fonte de informação poderia ser usada por uma ferramenta como uma base de referência e, por outra, como uma base de fontes. Isso dependeria do tipo de dado recuperado a partir dela por cada ferramenta.

Tanto no que tange as fontes de informações, quanto no que se refere ao modelo usado para recuperá-las, sem dúvida há divergências entre o ambiente tradicional de uma biblioteca e o pouco usual de uma coleção pessoal. Porém, ao se avaliar o catálogo resultante desse processo automatizado, segundo o grau de adequação aos elementos do ISBD, percebe-se que somente os usuários mais preciosistas atentariam para os poucos elementos faltantes. O que em nada desabona a utilidade do processo ou das ferramentas que o executam.

Contudo, para encerrar este estudo, segue uma lista com recomendações aos

desenvolvedores desse tipo de *software*:

a) implementar os elementos faltantes do ISBD;

b) salvar sempre os novos metadados descritivos diretamente nos arquivos;

c) integrar a ferramenta às bases de dados de referência on-line.

O que motiva a primeira recomendação é ampliar a aceitação desse tipo de ferramenta entre bibliotecários consultores, além de contemplar o preciosismo dos usuários que hoje inserem essas informações no campo destinado aos comentários sobre o documento. A segunda recomendação contribui com a segurança dos dados da coleção. Pois se algo acontecer com a base de dados local da ferramenta, todos os principais dados descritivos poderiam ser recuperados dos próprios arquivos de mídia, sem ter que recorrer a fontes de informação da Internet. E, a terceira recomendação visa permitir que o usuário da ferramenta consiga, através dela, alimentar as bases de dados on-line com informações sobre os itens de sua coleção. Ampliando, desse modo, a abrangência da catalogação social.

REFERÊNCIAS

10-FOOT user interface. In: **XBMC Wiki**. 22 Mar. 2012. Disponível em: <http://wiki.xbmc.org/index.php?title=10-foot_user_interface>. Acesso em: 25 abr. 2012.

ALBA, A.; BHAGWAN, V.; GRANDISON, T.. Accessing the Deep Web: when good ideas go bad. In: Association for Computing Machinery (ACM). **OOPSLA Companion 8**. ACM: New York, 2008, p. 815-818.

AMERICAN NATIONAL STANDARDS INSTITUTE (ANSI). Standards Planning and Requirements Committee (SPARC). Interim Report: ANSI/X3/SPARC study group on data base management systems.**FDT:** bulletin of ACM SIGMOD, New York, v.7, n.7, p.1-40, 1975.

ANOBII. **Anobii**: Where Books and People Meet. Disponível em: <http://www.anobii.com/about>. Acesso em: 12 nov. 2011.

ARMS, W. Y.. **Digital Libraries**. Cambridge: MIT, 2000. Disponível em: <http://www.cs.cornell.edu/wya/diglib/index.html>. Acesso em: 3 abr. 2012.

BALDRIDGE, A.. **Organize Your Digital Life**: how to store your photographs, music, videos, and personal documents in a digital world. Washington D.C.: National Geographic, 2009.

BEAGRIE, N.. **Plenty of Room at the Bottom?** Personal Digital Libraries and Collections. Dlib Magazine, v.11, n.6, Jun. 2005. Disponível em: <http://www.dlib.org/dlib/june05/beagrie/06beagrie.html>.Acesso em: 14 nov. 2011.

BREEDING, M.. Understanding the Protocol for Metadata Harvesting of the Open Archives Initiative. **Computers in Libraries**, Nashville, v.22, n.8, p. 24-29, Sep. 2002.

BRYSON, B. **At Home**: A Short History of Private Life. New York: Doubleday, 2010.

BUSH, V.. As We May Think. In: Trend, D. (Org.). **Reading Digital Culture**. Oxford: Blackwell, 2001. p.9-13.

CAPLAN, P.. **Metadata Fundamentals for all Librarians**. American Library Association: Chicago, 2011.

CASTRO, F. F. de; SANTOS, P. L. V. A. da C.. Os Metadados como Instrumentos Tecnológicos na Padronização e Potencialização dos Recursos Informacionais no Âmbito das Bibliotecas Digitais na Era da Web Semântica. **Informação & Sociedade:** estudos, João Pessoa, v. 17, n. 2, p. 13-19, maio/ago. 2007.

COX, M.; MULDER, E.; TADIC, L.. **Descriptive Metadata for Television**: an end-to-end introduction. Waltham: Focal, 2006.

CUNHA, M. B. da. Bases de Dados no Brasil: um potencial inexplorado. **Ciência da Informação**, v. 18, n. 1, p. 45-57, jan./jun. 1989.

DATE, C. J.. **Introdução a Sistemas de Bancos de Dados**. Rio de Janeiro: Elsevier, 2004.

FEATURES and supported codecs. In: **XBMC Wiki**. 28 Mar. 2012. Disponível em: <http://wiki.xbmc.org/index.php?title=Features_and_supported_codecs>. Acesso em: 25 abr. 2012.

GARDNER, S. A.. The Changing Landscape of Contemporary Cataloging. **Cataloging & Classification Quarterly**, Lincoln: Haworth, v.45, n. 4, p. 81-99, 2008. Disponível em: <http://digitalcommons.unl.edu/libraryscience/168>. Acesso em: 13 abr. 2012

GIUSTINI, D.; HOOKER, D.; CHO, A.. Social Cataloguing: an overview for health librarians. **Journal of the Canadian Health Libraries Association**, Thunder Bay, v.30, n.4, p.133-138, 2009.

GOMEZ, J.. **Print is Dead**: books in our digital age. New York: Macmillan, 2008.

GOYAL, K.. **About Calibre**. Nov. 2009. Disponível em: <http://calibre-ebook.com/about#history>. Acesso em: 1 jun. 2012.

GOYAL, K.. **Calibre User Manual**. 31 May 2012. Disponível em: <http://manual.calibre-ebook.com/index.html>. Acesso em: 1 jun. 2012.

GRACENOTE. **Gracenote Global Media Database.** Disponível em: <http://www. gracenote .com/products/global_media_database/>. Acesso em: 11 nov. 2011.

HIRWADE, M. A.; BHERWANI, M. T.. Facilitating Searches in Multiple Bibliographical Databases: metadata harvesting service providers. **Liber Quarterly**, The Hague, v. 19, n.2, p. 140-165, Dec. 2009.

HOWE, J.. **Crowdsourcing**: why the power of the crowd is driving the future of business. New York: Three Rivers, 2009.

INTERNET MOVIE DATABASE (IMDB). **Contributor Zone**. Disponível em: <http://www. imdb.com/czone/>. Acesso em: 10 nov. 2011.

JENKINS, C.. **Clementine Music Player**. 27 Dec. 2011. Disponível em: <http://www.clementine-player.org>. Acesso em: 2 jun. 2012.

KIM, H.; CHOO, C.-Y.; CHEN S-S.. Generating a Meta-DL by Federating Search on OAI and Non-OAI Servers. **Journal of Intelligent Information Systems**, Norwell, v. 34, n. 2, p. 177-191, Apr. 2010.

LAKATOS, E. M; MARCONI, M.. **Fundamentos de Metodologia Científica**. 3. ed. São Paulo: Atlas, 1991.

LEE, C. A. (Ed.). **I, Digital**: personal collections in the digital era. Chicago: SAA, 2011.

LIBRARYTHING. **About LibraryThing**. Disponível em: <http://www.librarything.com/about>. Acesso em: 11 nov. 2011.

MACEDO, N. D. de. **Iniciação à Pesquisa Bibliográfica**: guia do estudante para a fundamentação do trabalho de pesquisa. 2.ed. São Paulo: Loyola, 1996.

MARCONI, M. de A.; LAKATOS, E. M.. **Técnicas de Pesquisa**: planejamento e execução de pesquisas, amostragens e técnicas de pesquisa, elaboração, análise e interpretação dos dados. 6. ed. São Paulo: Atlas, 2006.

MARCONDES, C. H.. Metadados: descrição e recuperação de informação na Web. In: MARCONDES, C. H.; KURAMOTO, H.; TOUTAIN, L. B.; SAYÃO; L. (Orgs.). **Bibliotecas Digitais**: saberes e práticas. Brasília: IBICT, 2006. p. 95-111.

MATROSKA. **Matroska FAQ**. 9 Set. 2011. Disponível em: <http://www.matroska.org/technical/guides/faq/index.html>. Acesso em: 29 maio 2012.

McGREGOR, G.. Scrape Your Way to Happiness. **J-Source.ca**, Ottawa: CJP, 11 Jan 2012. Disponível em: <http://j-source.ca/article/web-scraping-how-journalists-get-their-own-data>. Acesso em: 20 maio 2012.

MUCHERONI, M. L.; SILVA, J. F. M. da. A Interoperabilidade dos Sistemas de Informação sob o Enfoque da Análise Sintática e Semântica de Dados na Web. **Ponto de Acesso**, Salvador, v. 5, n. 1, p. 3-18, 2011.

MUSICBRAINZ. **General FAQ**. Disponível em: <http://musicbrainz.org/doc/General_FAQ>. Acesso em: 13 nov. 2011.

NATIONAL INFORMATION STANDARDS ORGANIZATION (NISO). **Understanding Metadata**. Bethesda: NISO, 2004. Disponível em: <http://www.niso.org/publications/press/UnderstandingMetadata.pdf>. Acesso em: 4 abr 2012.

NGUYEN, D.. **The Bastards Book of Ruby**: a programming primer for counting and other unconventional tasks. New York: [s.n.], 2011. No prelo. Disponível em: <http://ruby.bastardsbook.com/>.Acesso em: 20 maio 2012.

PENNYCOOK, B.. The Impact of New Media on Music Creativity. In: WINGET, M.; ASPRAY, W. (Eds.). **Digital Media**: technological and social challenges of the Interactive world. Lanham: Scarecrow,2011. p. 187-212.

RHODES, B.; GOERZEN, J.. **Foundations of Python Network Programming**. Berkeley: Apress, 2010.

ROB, P.; CORONEL, C.. **Database Systems**: design, implementation, and management. 8. ed. Boston: Cengage Learning, 2009.

ROWLEY, J.. **A Biblioteca Eletrônica**. 2. ed. Brasília: Briquet de Lemos, 2002.

ROWLEY, J.; HARTLEY, R.. **Organizing Knowledge**: an introduction to managing access to information. 4. ed. Guildford: Ashgate, 2008.

SANTOS, V. dos; CANDELORO, R. J.. **Trabalhos Acadêmicos**: uma orientação para a pesquisa e normas técnicas. Porto Alegre: AGE, 2006.

SIMEONI, F. Servicing the Federation: the case for metadata harvesting. In: EUROPEAN CONFERENCE ON RESEARCH AND ADVANCED TECHNOLOGY FOR DIGITAL LIBRARIES, 8., 2004, Bath.**Proceeding...** Berlin: Springer, 2004. p. 389-399.

SPITERI, L. F.. Social Cataloguing Sites: features and implications for cataloguing practice and the public library catalogue. In: CONGRESO ISKO-ESPAÑA, 9., 2009, Valencia. **Nuevas Perspectivas para la Difusión y Organización del Conocimiento**. Valencia: ISKO, 2009. p. 40-60.

SPITERI, L. F.. Using Social Discovery Systems to Leverage User-Generated Metadata. **Bulletin of the American Society for Information Science and Technology**, Silver Spring, v. 37, n. 4, p. 27-29, Apr./May 2011.

SULEMAN, H. Parallelising Harvesting. In: INTERNATIONAL CONFERENCE ON THE ASIAN DIGITAL LIBRARY, 9., 2006, Kyoto, **Proceedings...** Berlin: Springer, 2006. p. 81-90.

TAG editor. In: **The Amarok Handbook**. 9 Dez. 2011. Disponível em: <http://userbase.kde.org/Amarok/Manual/Organization/TagEditor>. Acesso em: 30 abr. 2012

THE TELEVISION DATABASE. **Welcome to The Television Database Wiki**. Disponível em: <http://www.thetvdb.com/wiki/index.php/Main_Page>. Acesso em: 12 nov. 2011.

TOMAÉL, M.I. (Org). **Fontes de Informação na Internet**. Londrina: EDUEL, 2008.

VEITH, R.H.. Memex at 60: Internet or iPod? **Journal of the American Society for Information Science and Technology**, New York, v. 57, n. 9, p. 1233-1242, Jul. 2006.

WARD, P.; DAFOULAS, G.. **Database Management Systems**. London: Thomson Learning, 2006.