TIAGO RODRIGO MARÇAL MURAKAMI

TESAUROS E A WORLD WIDE WEB

Trabalho de Conclusão de Curso apresentado ao Departamento de Biblioteconomia e Documentação da Escola de Comunicações e Artes da Universidade de São Paulo como requisito parcial para a obtenção do título de Bacharel em Biblioteconomia e Documentação.

Orientadora: Profª Drª Sueli Mara Ferreira

São Paulo

2005

Autorizo:

[ ] divulgação do texto completo em bases de dados especializadas.

[ ] reprodução total ou parcial, por processos fotocopiadores, exclusivamente para fins acadêmicos e científicos.

Assinatura: _____

Data: _____

Murakami, Tiago Rodrigo Marçal

Tesauros e a World Wide Web / Tiago Rodrigo Marçal Murakami. – São Paulo: T.R.M. Murakami, 2005.
92 p.

Trabalho de Conclusão de Curso – Escola de Comunicações e Artes/USP, 2005.

Tesauros documentários
World Wide Web

Termos de Aprovação

Nome do Autor: Tiago Rodrigo Marçal Murakami

Título da monografia: Tesauros e a World Wide Web

Presidente da Banca: Profª Drª Sueli Mara Ferreira

Banca Examinadora:

Profª. Drª. Instituição: USP

Profº Dr. Instituição:

Aprovada em:

___/___/___

A Katian, Dai e Iuri ...

Agradecimentos

A todos os Professores do Departamento de Biblioteconomia e Documentação, por serem responsáveis pela minha formação, e especialmente às Professoras Doutoras Daisy Pires Noronha, Sueli Mara Ferreira e Marilda Lara, pela imprescindível ajuda no decorrer do trabalho.

A todos os amigos discentes da ECA/USP, pelo companheirismo e amizade.

Aos amigos discentes das Universidades Uni-Rio, UFMG, UnB, UFPE, UFRN, UFSC, UFF, UDESC e UFSCar que conheci em encontros estudantis pois me ajudaram a enxergar novas possibilidades da profissão e a conhecer outras realidades.

Ao pessoal da Biblioteca Jurídica do Banco Itaú S.A. pelo companheirismo e força, além da paciência por conviverem comigo diariamente.

Resumo

MURAKAMI. T. R. M. **Tesauros e a World Wide Web**. São Paulo, 2005. Monografia (Trabalho de Conclusão de Curso) – Curso de Biblioteconomia e Documentação, Escola de Comunicação e Artes, Universidade de São Paulo.

Os tesauros são ferramentas que estão ganhando crescente importância no contexto da Web. Para que isso seja possível, foi necessário adaptar os tesauros para as tecnologias e as funcionalidades da Web. O presente trabalho é um estudo exploratório que tem como objetivo identificar como os Tesauros Documentários estão sendo utilizados e/ou incorporados à nova dinâmica de gestão de informações na Web.

Abstract

MURAKAMI. T. R. M. **Thesauri and the World Wide Web**. São Paulo, 2005. Monografia (Trabalho de Conclusão de Curso) – Curso de Biblioteconomia e Documentação, Escola de Comunicação e Artes, Universidade de São Paulo.

Thesauri are tools that growing importance in Web context. For this, is necessary adapting the thesauri for Web technologies and functionalities. The present work is an exploratory study that aim identifies how the documentary thesauri are being utilized and/or incorporated for the management of information in the Web.

Sumário

Índice de Quadros

Índice de Figuras

Lista de Abreviaturas e siglas

ANSI – American National Standards Institute

API – Application Program Interfaces

ISO - International Organization for Standardization

KOS - Knowledge Organization Systems

LD – Linguagens Documentárias

NISO – National Information Standards Organization

OWL - OWL Web Ontology Language

RDF - Resource Description Framework

SKOS – Simple Knowledge Organization System

SRI – Sistemas de Recuperação da Informação

W3C – World Wide Web Consortium

WWW ou Web – World Wide Web

XML – eXtensible Markup Language

1 Introdução

Os Tesauros Documentários, segundo a norma norte americana ANSI/NISO Z39.19-2003, são:

> ”Vocabulários controlados organizados em uma ordem conhecida em que as relações de equivalência, homográficas, hierárquicas e associativas entre os termos são claramente exibidas e identificadas por indicadores padronizados de relacionamentos.” (ANSI/NISO Z39.19-2003)

Eles surgiram na década de 50 com o propósito de servir de ajuda para ampliar o vocabulário de indexadores e devido às combinadas pressões de surgimento de novas áreas de assuntos e coleções, de novos modelos no uso da informação e expansão de aplicações de armazenamento e de processamento e recuperação da informação em computadores, foram aperfeiçoados para promover o controle terminológico de sistemas de informação e se tornar uma estrutura conceitual de um determinado campo do conhecimento.

Desde então, são principalmente utilizados para promover o controle de vocabulário em sistemas de recuperação da informação (SRI). Para isso, são utilizados pelos indexadores no momento da indexação e devem ser disponibilizados para o usuário no momento da recuperação.

Com o surgimento e posterior desenvolvimento da World Wide Web (Web), os tesauros documentários começaram também ser utilizados nesse ambiente informacional.

A relação entre Tesauros Documentários e a Web é bilateral, na qual ambos ganham. No princípio dos anos 90, a Web foi utilizada para a distribuição de Tesauros e posteriormente, devido à migração de Sistemas de Recuperação da Informação para esse ambiente e a crescente demanda por organização da informação da Web, os tesauros começaram ser utilizados para a organização da própria Web. Mas para que pudessem ser utilizados na Web, os Tesauros Documentários precisaram passar por um processo de adaptação para esse ambiente.

Primeiramente, a adaptação foi tecnológica. Os Tesauros Documentários tiveram que ser representados em um formato tecnológico compatível com os padrões vigentes na Web. É

importante ressaltar que essa adaptação ainda continua, pois a Web ainda se mantém em pleno desenvolvimento e a representação de tesauros precisa acompanhá-la.

Outra adaptação necessária para a utilização dos Tesauros na Web é em relação às funções que os Tesauros exercem nela. Eles não poderão ser utilizados da mesma maneira que são utilizados nos SRI, isto é para promover somente o controle terminológico, conforme alerta Sajus (2002):

> "A função tesaural deverá ter um papel importante nas tecnologias de acesso à informação por conteúdo, contanto que não o reduzam às práticas tradicionais de indexação documentária. É ilusório imaginar o futuro de sistemas de informação, inclusive o futuro da Web, a partir de práticas concebidas para e pelos centros documentação" (SAJUS, 2003)

Isso acontece porque a Web é um sistema de comunicação aberto e heterogêneo e essa estrutura inviabiliza o uso de tesauro somente para o controle de vocabulário na maioria dos ambientes de informação presentes nela. Porém, as ricas estruturas conceituais e semânticas dos tesauros documentários estão sendo utilizadas para exercer em novas funções na Web. Porém, por ser uma utilização recente, não há estudos amplos sobre novas funções dos tesauros na Web. Levando isso em consideração, o nosso trabalho terá o seguinte objetivo:

1.1 Objetivo

Identificar como os Tesauros Documentários estão sendo utilizados e/ou incorporados à nova dinâmica de gestão de informações na Web, por meio de um estudo exploratório, para chegar a uma possível sistemática sobre o tema.

1.2 Metodologia

Para atingir o objetivo, faremos um estudo exploratório na literatura da área de Ciência da Informação e da Computação. Os estudos exploratórios, segundo Dencker & Viá (2001), são:

> "Investigações de pesquisa empírica que têm por finalidade formular ou esclarecer questões para desenvolver hipóteses. O estudo exploratório

> aumenta a familiaridade do pesquisador com o fenômeno ou com o ambiente que pretende investigar, servindo de base para uma pesquisa futura mais precisa. São também utilizados para esclarecer ou modificar conceitos. As descrições, nesse caso, tanto podem ser qualitativas quanto quantitativas. Os métodos de coleta de dados também podem variar da pesquisa bibliográfica e documental ao uso de questionário, entrevista ou observação." (DENCKER & VIA, 2001)

O método de coleta de dados escolhido é o "*documentary research*" ou pesquisa documentária, que é descrito por Busha & Harter (1980) da seguinte maneira: *"the generic term documentary research is used here to refer to inquires into the printed tools of librarianship – book, journal, and indexes*."[1]. Ela será feita nas seguintes fontes: as bases de dados Web of Science[2], E-LIS[3], ERIC[4], LISA[5], CiteSeer[6], Dedalus[7], Metasearch da Universidade de Hanover[8], Elsevier Science Direct[9], Archive Ouverte en Sciences de l'Information et de la Communication[10], Portal ACM[11], revistas cientificas nacionais da área[12], ferramentas de busca na Internet[13] e o acervo da Biblioteca da Escola de Comunicações e Artes[14] da Universidade de São Paulo.

---

[1] Tradução nossa: "O genérico termo pesquisa documentária é usado aqui se referindo a perguntas às ferramentas impressas da biblioteconomia – livros, periódicos e índices."

[2] acesso pelo SIBi/USP.

[3] http://eprints.rclis.org

[4] http://www.eric.ed.gov/

[5] acesso pelo SIBi/USP.

[6] http://citeseer.ist.psu.edu/

[7] acesso pelo SIBi/USP.

[8] http://www.tib.uni-hannover.de/en/digital_library/metasearch/

[9] http://www.sciencedirect.com

[10] http://archivesic.ccsd.cnrs.fr/

[11] http://portal.acm.org/portal.cfm

[12] Periódicos disponíveis na Biblioteca da ECA/USP.

[13] Google - http://www.google.com.br

Google FR - http://www.google.fr

Google Scholar – http://scholar.google.com

[14] http://www.rebeca.eca.usp.br/

As estratégias de busca utilizadas são os termos thesaurus, tesauro, tesauros, thesauri e Web no período de 1997 a 2004, com preferência aos resultados nos idiomas Português, Inglês, Francês e Espanhol. Após observar os resultados das primeiras pesquisas, decidimos modificar a estratégia para incluir também, os termos "*Knowledge Organization Systems*" e Linguagens Documentárias.

Os resultados obtidos serão sistematizados sob dois focos:

1. Representação tecnológica dos tesauros documentários na Web;
2. Emprego dos tesauros documentários em várias etapas da gestão da informação em ambientes informacionais da Web.

2 O Tesauro Documentário

O termo “Thesaurus” é a forma em Latim da palavra Grega “thesauros”, que significava originalmente “estoque de tesouros” ou em inglês “*treasure store*”. No século 16, ele começou a ser usado como sinônimo para dicionário (um estoque de tesouros de palavras), mas posteriormente isso caiu em desuso. Peter Mark Roget ressuscitou o termo em 1852 para o titulo de seus dicionários de sinônimos. A proposta desse trabalho era dar ao usuário a escolha entre termos similares quando a primeira diretamente delas não dá a impressão de se ajustar perfeitamente. Cem anos depois, no começo dos anos 50, a palavra “thesaurus” começou a ser empregada também como o nome para uma lista de palavras, mas com o objetivo exatamente oposto ao de Roget: determinar o uso de somente um termo (um “descritor”) para um conceito que pode ter sinônimos. A similaridade entre o *Roget´s Thesaurus* e o tesauro para indexação e recuperação da informação é que ambos listam os termos relacionados hierarquicamente ou associativamente para descritores, somado aos sinônimos. (ANSI/NISO Z39.19-2003)

O primeiro tesauro documentário moderno surgiu em 1959[15] como uma ajuda para os indexadores estenderem seus vocabulários, porém era usado de modo secundário (mais propriamente sugestivo que prescritível). O tesauro foi desenvolvido por um período de 15 anos sob combinadas pressões de crescimento rápido, conjunto de novas áreas de assuntos e coleções, de novos modelos no uso da informação e expansão de aplicações de armazenamento, processamento e recuperação da informação em computadores. Embora o primeiro grupo de especialistas em informação escolhessem abandonar os princípios ortodoxos de recuperação aplicados por bibliotecas, particularmente por esquemas de classificação bibliográficos, um segundo grupo liderado por Calvin Mooers e Charles Bernier defenderam o valor do controle terminológico e estrutura conceitual; o pensamento e trabalho do segundo grupo lideraram o desenvolvimento do tesauro como conhecemos hoje. A integração de relacionamentos durante o desenvolvimento do tesauro foi sugerida por Bernier

[15] O primeiro tesauro da era moderna é considerado o desenvolvido em 1959 pelos indexadores do *E.I. Dupont de Nemours Engineering Department*. Esse tesauro, porém, não foi publicado ou vastamente distribuído. O *Chemical engineering thesaurus*, um derivado direto da ferramenta de *Dupont*, foi publicado em 1961 (Hudon 2003)

em 1957 e após isso, muita ênfase foi dada a esse componente, pois essa estrutura o distingue dos outros tipos de linguagens controladas de indexação. (HUDON, 2003)

Hoje os tesauros documentários são construídos com base em normas criadas por entidades internacionais e nacionais em alguns países. A principal é a *Guidelines for the Establishment and Development of Monolingual Thesauri ou* **ISO 2788***,* de 1986, publicada pela *International Organization for Standardization (ISO)*, entidade normalizadora internacional. Essa norma deu origem a normas em diversos países. Em nosso trabalho, utilizaremos como referência apenas duas:

1. **Brasil:** A versão em português da ISO 2788: Diretrizes para o estabelecimento e desenvolvimento de Tesauros Monolíngües, publicada pelo Senai/IBICT em 1993.

2. **Estados Unidos:** A norma dos elaborada pela *American National Standards Institute / National Information Standards Organization (U.S.): Guidelines for the Construction, Format, and Management of Monolingual* ***Thesauri*** de 2003 ou **ANSI/NISO Z39.19-2003.** Essa norma foi baseada em diversas normas americanas e internacionais sobre a criação de tesauros, incluindo a ISO 2788. Atualmente, a ANSI/NISO Z39.19 está em processo de revisão: em abril/2005, foi disponibilizado um esboço dessa revisão para discussão que utilizaremos em nosso trabalho apenas como base de comparação. Nos reportaremos a essa revisão da seguinte maneira: **ANSI/NISO Z39.19-200X.** Como informação complementar, faremos uma pequena comparação entre a norma e a revisão que será incluída no Apêndice A.

No presente estudo, daremos uma maior importância à norma **ANSI/NISO Z39.19-2003** por que a consideramos mais atual em relação à norma ISO 2778 de 1986 e mais estável em relação à sua revisão, a **ANSI/NISO Z39.19-200X**.

O tesauro documentário é definido pela norma ANSI/NISO Z39.19-2003 da seguinte maneira:

> ”Vocabulários controlados organizados em uma ordem conhecida em que as relações de equivalência, homográficas, hierárquicas e associativas entre os

> termos são claramente exibidas e identificadas por indicadores padronizados de relacionamentos." [16]

O tesauro documentário é composto de descritores (que podem apresentar qualificadores parentéticos quando necessário), notas de escopo e notas de escopo recíprocas e relacionamentos entre os descritores.

Os relacionamentos permitidos são de três tipos:

a) Relacionamento de equivalência;
b) Relacionamento hierárquico;
c) Relacionamento associativo.

Para uma descrição mais completa dos elementos e relacionamentos presentes nos tesauros documentários, fizemos uma tradução, com propósito acadêmico, da norma ANSI/NISO Z39.19-2003 e incluímos no Anexo A.

Os tesauros são identificados na literatura como pertencentes a diversas categorias (ou famílias), durante o nosso trabalho, identificamos as seguintes: Linguagens Documentárias (ou LD), *Knowledge Organization Systems, ou KOS* (que pode ser traduzido como Sistemas de organização do conhecimento) e *Ressources terminologiques ou ontologiques, ou RTO* (que pode ser traduzido como Recursos terminológicos ou ontológicos). A diferença está na função para que o tesauro documentário é utilizado, como podemos observar através das definições dadas:

Linguagem documentária, segundo Gadin (apud CINTRA ET AL., 2002, p.35) "é um conjunto de termos providos ou não de regras sintáticas, utilizado para representar conteúdos de documentos técnicos-cientificos, com fins de classificação ou busca de informação".

Já "*Knowledge Organization Systems"* (KOS) apresentam uma finalidade mais ampla:

> "Sistemas de Organização do Conhecimento (*Knowledge Organization Systems*) podem abranger tesauros e outras listas controladas de palavras-

[16] Tradução nossa a partir do texto: "A controlled vocabulary arranged in a known order in which equivalence, homographic, hierarchical, and associative relationships among terms are clearly displayed and identified by standardized relationship indicators, which must be employed reciprocally." (ANSI/NISO Z39.19, 2003, p.1)

> chave, ontologias, sistemas de classificação, similares a clusterização, taxonomias, dicionários geográficos, bases de dados lexicais, mapas conceituais/espaciais, mapas semânticos, etc. Esses esquemas permitem a estruturação e gerenciamento do conhecimento, processamento de dados baseado em conhecimento e acesso sistemático para estruturas de conhecimento em coleções individuais e bibliotecas digitais. Usado como serviços interativos de informação na Internet eles tem um potencial crescente para suportar a descrição, descoberta e recuperação de recursos heterogêneos de informação e contribuir para uma infra-estrutura de descoberta de recursos completa" (HILL, 2001)

E Ressources terminologiques ou ontologiques (RTO) também tem um propósito um pouco diferente:

> "A disponibilização dos usuários de documentos em formato eletrônico representa hoje uma verdadeira aposta cientifica. Essa aposta, associada à demanda social em ligação com o tratamento de dados textuais contidos nesses documentos, faz emergir uma nova problemática, visando a modelizar o conteúdo dos documentos selecionados ou formatando uma coleção sobre a forma de entrelaçamento de termos para permitir um melhor acesso ao conhecimento. Esses modelos ou representações podem ser entre outros tesauros, terminologias, linguagens documentarias, índices ou ontologias. Nós os chamaremos de recursos terminológicos ou ontológicos (RTO)." (AUSSENAC-GILLES & CONDAMINES, 2004)

2.1 Construção de tesauros

Identificamos na literatura que os tesauros documentários poderão ser construídos de duas maneiras: por especialistas e automaticamente. O foco do presente trabalho é a utilização de tesauros construídos apenas por especialistas, mas os trabalhos que utilizam tesauros construídos automaticamente devem ser considerados pelo principio de automatização empregado.

2.1.1 Construção de tesauro por especialistas

Os tesauros construídos por especialistas seguem normas como a ISO 2788 e a ANSI/NISO Z39.19-2003. Segundo Batty (1998), a abrangência desses tesauros se dá em dois níveis:

1. **Áreas relacionadas** (*Broad areas*) – quando o seu conteúdo abrange várias áreas dentro de uma mesma especialidade ou entre áreas correlatas, como por exemplo: a Lista de Cabeçalhos de Assuntos Médicos (*Medical Subject Headings - MeSH*) da Biblioteca Nacional de Medicina dos EUA (*U.S. National Library of Medicine*), o Tesauro de Engenharia e Termos Científicos (*Thesaurus of Engineering and Scientific Terms - TEST*) originalmente do *Engineers Joint Council*, ou o Tesauro de Arte e Arquitetura *(Art and Architecture Thesaurus - AAT)* desenvolvido pela *Getty Trust*; e

2. **Áreas específicas** (*Specific areas*) – quando atende a uma área especializada, como por exemplo: o *Transportation Research Thesaurus* (TRT) administrado pela *Transportation Research Board* da *National Research Council* ou o ERIC Thesaurus na educação.

Bourigault et al (2004) e Aussenac-Gilles & Condamines (2004), complementam, ao afirmar que existem também tesauros construídos sobre o corpus de coleções específicas, que têm como abrangência as próprias coleções.

2.1.2 Criação automática de tesauros

Tesauros são criados automaticamente com base em uma coleção de documentos eletrônicos, com o objetivo de melhorar a recuperação da informação em sistemas de texto completo. Eles são utilizados pela área de computação, para finalidades específicas. Esses tesauros criados por máquinas não são idênticos aos criados por especialistas, pois utilizam apenas algumas relações entre termos (normalmente um ou dois tipos de relações) e tem como abrangência a coleção de documentos sobre a qual foram construídos.

A criação de tesauros por máquinas não é o foco deste trabalho, mas indicaremos algumas formas de construção e trabalhos relacionados: **Construção por Similaridade *(Similarity Thesauri):*** ZAZO ET AL. (2004) e **Construção por Associação (*Association Thesaurus*):** KAJI ET AL (2000) e YANG & KUK (2003)

3 Tesauros e World Wide Web

A utilização de tesauros na Web é crescente. O tesauro está sendo utilizado em ambientes em que normalmente não era utilizado, devido às necessidades de organização de vocabulário criadas pelo crescimento na Web. A Web, de um modo bem resumido, é um sistema aberto baseado em padrões tecnológicos de comunicações entre máquinas (ex. http, ftp, entre outros) e entre humanos (ex. html, dhtml, entre outros) que está em constante evolução. Esse processo de evolução é feito de modo descentralizado.

A representação dos tesauros eletronicamente tornou possível o tesauro ser utilizado na Web. Esse caminho se deu da seguinte maneira:

As primeiras vantagens de representar um tesauro eletronicamente foram listadas por Davies (1995 apud HUDON, 2003):

- A redução dos custos de comercialização como produção, armazenamento e distribuição de produtos impressos;
- Implica em aumentar o uso de tesauros devido à diminuição dos custos de aquisição e extensão da disponibilidade e;
- Facilidade de atualização assegurando que os usuários poderão sempre ter acesso às últimas adições e modificações.

Essas vantagens não estão propriamente ligadas a Web, mas a Web é um potencial canal de distribuição. Já Cueva Martín (1999) destaca que a Internet, sobretudo a World Wide Web, oferece vantagens e novas possibilidades de desenvolvimento e acesso a tesauros online, entre as quais ele destaca:

- É um meio idôneo para desenvolver a estrutura Hypertextual da rede semântica de um tesauro, com links entre os termos que mantém uma relação de equivalência, hierarquia ou associativa e entre as diferentes partes do tesauro (alfabética, hierárquica e permutada) e a possibilidade de navegar entre elas. Também podem se estabelecer links com imagens e sons.

- Neste meio, pode se simplificar muito a estrutura dos tesauros e facilitar seu manejo com interfaces adequadas, em relação às versões impressas.

- Redução significativa dos custos de atualização. Pode ser uma alternativa as edições impressas, habitualmente caras.

- Pode contribuir para o desenvolvimento de tesauros multilíngües e multidisciplinares com equipes de trabalho de organizações de distintos países.

- Permite dispor de ferramentas terminológicas, de acesso universal, de ajuda a recuperação em distintas bases de dados e como fonte para estabelecer bases de conhecimento.

No entender de Cueva Martín, a Web facilita o desenvolvimento dos tesauros, mas a Web continua como um canal de distribuição e desenvolvimento de tesauros.

Isso muda na visão de Shiri & Revie (2000), pois eles afirmam que as razões que levam ao desenvolvimento de tesauros na Web estão intimamente ligadas às necessidades da World Wide Web:

- O colossal crescimento dos recursos informacionais demanda uma melhor identificação dos seus assuntos;

- A migração de tradicionais recursos informacionais para a Web clama por mais consistentes aproximações por assunto;

- Uma urgente necessidade para descrição de recursos e descoberta do reuso direto de ferramentas de gerenciamento da informação como os vocabulários controlados;

- Problemas associados com a qualidade da informação não-estruturada recuperada na Web;

- A necessidade de prover aos usuários estruturas de conhecimento como os tesauros para rápido e fácil acesso a informação melhor organizada.

Com isso, o tesauro que apenas utilizava a Web como um meio para seu próprio desenvolvimento, começa a ser utilizado para melhorar os processos de organização da Web. Para isso ele deve ser adaptado à tecnologia da web, conforme estudaremos a seguir.

3.1 Adaptação tecnológica dos tesauros documentários para a Web

O processo de adaptação dos tesauros para a World Wide Web foi gradual e acompanha o próprio desenvolvimento da Web. Eles foram adaptados em relação às tecnologias de representação de tesauros e a forma de acesso e visualização, conforme veremos a seguir:

3.1.1 Tecnologias de representação de tesauros na Web

Shiri & Revie (2001) fizeram uma listagem dos tipos de tecnologias de representação dos tesauros presentes na Web:

1. Tesauro em um formato de texto estático simples;
2. Tesauro no formato HTML mas ainda estático, sem o uso efetivo de hyperlinks;
3. Tesauro no formato HTML com hyperlinks completamente navegáveis;
4. Tesauro com interface gráfica e visual avançada;
5. Tesauro em formato XML.

Essa listagem apresenta a provável seqüência de adaptação dos tesauros para a web:

O **formato de texto estático simples** é a simples disponibilização de tesauros em documentos como txt, rtf, doc, e outros tipos de arquivos criados por editores de texto na Web. Esses arquivos são representações idênticas dos tesauros em papel e servem apenas para consulta, download ou para serem impressos.

Os **Tesauros no formato HTML, mas ainda estático sem o uso efetivo de hyperlinks**, é a representação do formato anterior na linguagem de marcação HTML, para visualização em browsers Web. Com isso, ele serve ainda apenas para consulta e impressão.

Os **Tesauros no formato HTML com hyperlinks completamente navegáveis** apresentam uma evolução em relação aos formatos anteriores. Ele utiliza os hyperlinks, ligações criadas

na Web, que permitem navegar entre textos ou dentro do próprio texto de forma rápida. Com isso, os tesauros se tornam navegáveis e mais fáceis de serem consultados.

Os **Tesauros com interface gráfica e visual avançada** são um aperfeiçoamento do formato anterior, que permite uma apresentação mais fácil de navegação e avanços gráficos como a visualização de mapas de redes de relacionamentos entre termos em três dimensões. Para tanto, utilizam tecnologias complementares como o Java.

Os **Tesauros no formato XML** apresentam a inovação de serem "manipuláveis por computadores". A linguagem *XML - eXtensible Markup Language*, ou Linguagem de Marcação Estendida, pode ser definida como um subconjunto da linguagem *SGML - Standard Generalized Markup Language*, que permite a criação de uma marcação própria com intuito de especificar idéias e compartilhá-las na grande rede, sendo controlada pelo consórcio *W3C - World Wide Web Consortium.* (FURGERI, 2001). Uma das características do *XML* é o fato de ser considerada uma **meta-linguagem**, o que significa que ela provê recursos para a definição de gramáticas que caracterizam linguagens para classes de documentos específicos, com conjunto de elementos, atributos e regras de composição bem determinadas. Porém, o uso da linguagem XML para a criação de tesauros apresenta o problema da limitação sintática do XML que não permite uma maior automatização dos processos. Como possibilidade de resolução desse problema, estão sendo desenvolvidos padrões de representação de tesauros usando uma tecnologia derivada do XML, o *Resource Description Framework* (ou simplesmente **RDF**).

O *Resource Description Framework (RDF)* é definido pela W3C como:

> "O *Resource Description Framework (RDF)* integra uma variedade de aplicações que vai desde catálogos de bibliotecas e diretórios mundiais para distribuição e agregação de notícias, software para coleções pessoais de música, fotos e eventos usando XML como uma sintaxe de intercâmbio. A especificação RDF provê um sistema de **ontologia leve** para suportar a troca de conhecimento na Web"

O uso do *Resource Description Framework (RDF)*, que foi proposto pela **World Wide Web Consortium (W3C)** pode prover a base para a interoperabilidade entre tesauros. O conceito *"RDF Namespace"* permite o uso controlado de sistemas de vocabulário distribuídos e também

provê uma sintaxe (XML) para exportar dados de vocabulário controlado com outras aplicações e serviços (KOCH, 1999 apud SHIRI & REVIE, 2001)

A representação de tesauros em RDF ainda é muito recente, e ainda há muito espaço para discussão sobre o tema. Miles & Matthews (2001) fizeram um resumo das iniciativas até então existentes de utilização de tesauros em RDF no documento "*Review of RDF Thesaurus Work*" para o projeto *SWAD-Europe (Semantic Web Advanced Development for Europe)* e desse trabalho surgiu o SKOS (*Simple Knowledge Organisation System*), que é um padrão de representação de tesauros usando o RDF. Atualmente o padrão se encontra na sua fase 2, mas ainda não se tornou especificação da W3C (está em "*W3C Working Draft*", ou esboço de trabalho da W3C).

3.1.2 Formas de acesso e apresentação dos tesauros

As formas de acesso e apresentação de tesauros na Web são influenciados pela tecnologia de representação de tesauros na Web, mas influenciam diretamente a forma de utilização dos tesauros.

O acesso a tesauros na Web, segundo Masse & Ménille (2004) é feito por meio de:

1. Download do arquivo, ou
2. Acesso pela Web (URL): Utilizando a navegação ou softwares baseados na Web.

E completa Johnson (2004): os tesauros podem ser acessados através aplicações especializadas, baixadas e instaladas no micro, que apenas se comunica com a Web.

O tesauros acesso a tesauros para download, torna o tesauros disponível apenas para consulta e impressão. Já o acesso pela Web, possibilita o acesso para consulta de modo mais dinâmico (tanto para indexadores, como para usuários finais) e permite uma melhor apresentação dos tesauros.

E o acesso a tesauros por meio de aplicações especializadas, permite uma maior participação do tesauro no Sistema de Recuperação da informação, pois esse modo de acesso possibilita uma maior integração entre o Tesauro e o SRI disponível na Web.

Já as formas de visualização de tesauros foram classificadas por Craven (2004) da seguinte maneira:

Interface de busca *(Search interface)*	As interfaces de busca permitem ao usuário buscar os termos através de consultas[17] ao tesauro.
Resultados de consultas *(Query results)*	Os tesauros são exibidos ao usuário no momento da exibição dos resultados.
Lista de termos *(Term list)*	Os tesauros são exibidos em uma lista simples de termos, normalmente em ordem alfabético-numérica.
Detalhes dos termos *(Term details)*	Detalhes dos termos como notas de escopo são exibidos durante a navegação.
Exibição hierárquica *(Hierarchical displays)*	Exibição do tesauro no modo hierárquico, permitindo a navegação entre: Termo Geral (TG) e Termo Específico (TE)
Exibição classificada *(Classified displays)*	A exibição classificada é utilizada quando os termos do tesauro recebem algum tipo de classificação.
Outros modos de exibição *(Other displays)*	Outros tipos de apresentações como apresentações expandíveis, apresentação do KWIC (Keyword in Context) e apresentação do KWOC (Keyword out of context)
Múltiplos modos de exibição *(Multiple displays)*	Modos de exibição que usam uma ou mais características apresentadas anteriormente. Modo mais comum entre os tesauros.
Navegação *(Navigation)*	Navegação pelo tesauro através de forma de apresentação gráfica.

Quadro 1 : Thesaurus displays on the Web

Fonte: CRAVEN (2004)

[17] Utilizando um formulário de busca.

As formas de visualização impactam diretamente no modo de utilização e devem ser utilizadas conforme a necessidade do Ambiente de informação. Outros modos de apresentação de tesauros na Web são apresentados no capítulo 9 da ANSI/NISO Z39.19-200X, que traz um sub-capítulo com considerações especiais de como apresentar de tesauros em browser Web, usando tecnologias da Web (capitulo 9.4.3 – *Web Format – Special Considerations*).

3.2 Utilização dos tesauros na Web

Para estudarmos como os tesauros estão sendo utilizados na Web, optamos por sistematizar a literatura sob três focos:

1. Funções dos tesauros na Web;
2. Técnicas para a utilização de tesauros na Web;
3. Ambientes informacionais em que os tesauros são utilizados.

3.2.1 Funções dos tesauros na Web.

A norma ANSI/NISO Z39.19-2003 descreve 4 propósitos para os tesauros:

1. **Tradução:** Para prover um modo para traduzir a linguagem natural dos autores, indexadores e usuários para um vocabulário controlado usado para indexação e recuperação.
2. **Consistência**: Para promover consistência na designação de termos de indexação.
3. **Indicação de Relacionamentos**: Para indicar relacionamentos semânticos entre termos.
4. **Recuperação:** Para servir como uma ajuda na busca e recuperação de documentos.

E a revisão ANSI/NISO Z39.19-200X inclui além desses:

5. **Nome e navegação:** Provê hierarquias claras e consistentes em um sistema de navegação para ajudar usuários a localizar objetos de conteúdo desejados.

Esses propósitos indicam somente os objetivos dos tesauros de modo geral. A literatura da área demonstra funções mais específicas para tesauros. Compilamos aqui a visão de vários autores:

Soergel (1997) lista as possíveis funções dos tesauros da seguinte maneira:

1. Prover um mapa semântico para campos individuais e relacionamentos entre e sobre *(across)* campos;
2. Melhora a comunicação em geral - Suporte para aprendizado e assimilarização da informação;
3. Provê base conceitual para o design (planejamento) de boa pesquisa e implementação e auxilia pesquisadores com o problema da clarificação;
4. Provê classificação para ação;
5. Suporte significativo, apresentação bem estruturada da informação;
6. Base conceitual para sistemas baseados em conhecimento;
7. Suporte para a recuperação da informação.

Sajus (2002) faz também uma lista não exaustiva das possíveis funções dos tesauros, ampliando essas funções para uma utilização mais automatizada:

- indexação documentária semi-automática;
- Gestão eletrônica de dicionários;
- Questões em linguagem natural de documentos pouco ou não estruturados;
- Classificação automática;
- Tradução assistida por computador;
- Ajuda a leitura rápida (“resumo automático”);
- Análise do discurso assistida por computador;
- Correção de texto assistida por computador;
- Geração automática ou semi-automática de texto;
- Disseminação seletiva da informação;
- Representação não textual de dados textuais.

As funções descritas acima não são necessariamente executadas no ambiente da Web, mas algumas podem ser adaptadas a ele.

Mais especificamente no ambiente web, Clarke & Yancey (2001) descreve possibilidades de aplicação utilizando sistemas automatizados, da seguinte maneira:

- Com um vocabulário controlado, ferramentas de indexação automática e classificadores têm um ponto de entrada em que podem analisar textos. Começando com uma lista de termos controlados e relacionamentos, como as que existem em um tesauro, filtram palavras irrelevantes (ruído) que poderiam normalmente ser encontradas em resultados de busca por palavras-chave.

- Por meio de associação de todos os conceitos sinônimos em um cluster de um único conceito controlado, resultados consistentes de busca podem ser acessados de forma indiferente pelo formulário do texto ou pela escolha de termos específicos através de qualquer página individual.

- Relacionamentos entre termos em um vocabulário controlado, como os relacionamentos hierárquicos, podem ser usados para auto expandir resultados de busca conforme a necessidade. Outros tipos de relacionamentos podem guiar usuários para conceitos que podem ser de interesse.

- Termos de vocabulário controlado podem ter peso mais forte que o texto em linguagem natural no sistema de busca, melhorando a precisão.

- Quando sistemas de busca inteligente estão analisando consultas de busca, eles podem fazer uso de relacionamentos entre termos, definições, e outros ricos atributos lingüísticos e semânticos de um vocabulário controlado.

E Soergel (2002) adaptou as funções já listadas por ele para o contexto das Bibliotecas Digitais da seguinte maneira (grifo do autor):

- **Suportar aprendizado e assimilização da informação.**

    - **Suportar aprendizado** sobre qualquer tópico **ao prover para** o aprendiz um **coerente framework conceitual apropriado para a sua idade.**

- **Aprendizado como recuperação da informação.** Framework conceitual para melhorar as perguntas feitas no sistema.
- **Auxilia leitores** no entendimento do texto.

- **Auxilia pesquisadores e usuários com o problema da clarificação:**
  - Provê a base conceitual para o design (planejamento) de uma boa pesquisa e implementação e para boa formulação de consultas. Inclui ajuda com:
  - **explorar o contexto conceitual de uma pesquisa ou problema prático –** um estudo, política, plano ou projeto de implementação e com **estruturação do problema**.
  - **Exemplos de funções específicas:**
  - **Apresenta os assuntos em um campo ou uma área de aplicação em um framework coerente.**

- **Auxilia na solução de problemas:** Auxilia na exploração das dimensões de um problema e aspectos a serem considerados na sua solução; provê uma classificação de aproximações para solucionar problemas específicos. Provê classificação e **definição consistente de variáveis para pesquisa / de critérios de avaliação para problemas práticos,** então melhorar a comparatibilidade da pesquisa e avaliação de resultados e torna a pesquisa mais cumulativa.

- **Suporta recuperação da informação:**
  - Provê **suporte baseado em conhecimento para buscas de usuários finais**. Suporta busca em múltiplas linguagens; buscas em texto livre; buscas em múltiplas bases de dados usando diferentes linguagens de indexação.
  - **Extração das necessidades dos usuários** por meio de uma série de menus baseados em arvore de busca, ou por meio de guiar na análise conceitual de um tópico de busca (questões baseadas em uma estrutura de faceta, apresentação de um segmento de uma hierarquia de conceitos para cada faceta aplicável).

- **Navegar a estrutura de classificação** para identificar conceitos úteis para a busca no nível de especificidade desejado. Navegar uma coleção, como um diretório de assunto.

- **Mapear os termos de consulta dos usuários para descritores** usados na base de dados **ou para as múltiplas expressões da linguagem natural** para serem usadas para busca em texto livre.

- **Busca inclusiva** (expandida hierarquicamente)

- **Melhorar algoritmos de rankeamento** com base em conceitos e relacionamentos entre termos.

- **Buscar múltiplas bases de dados** por meio do mapeamento dos termos de consulta dos usuários para descritores usados em cada base de dados, ou mapeamento de descritores de uma para outras bases de dados; linguagem de busca comum.

- **Suporta apresentação da informação**, especialmente apresentação de resultados de busca:

    - **Organização de unidades por significados** (registros de documentos, parágrafos, dados de propriedades de uma dada substancia encontrada a partir de diversas bases de dados), incluindo clusterização baseada em conhecimento de registros recuperados.

    - Isto suporta **a exploração de um amplo conjunto de recuperações e**, por extensão, **exploração do conteúdo de coleções inteiras** ou subcoleções**.**

    - **Organização da informação por significados em um registro** (por exemplo, ordenar os descritores encontrados)

- **Provê uma ferramenta para indexação.**

- **Controle de vocabulário.**

- **Indexação centrada no usuário** (orientada a consulta, orientada a problemas).
- Indexar **diversas bases de dados** em um campo com uma **linguagem comum de indexação** e compartilhar os resultados da indexação para reduzir completamente os esforços de indexação.
- **Mapeamento de descritores de indexação de um sistema para outro**
- **Facilitar a combinação de múltiplas bases de dados ou acesso unificado a múltiplas bases de dados** por meio de:
- **Mapeamento dos termos de consulta dos usuários** para os descritores usados em cada uma das bases de dados;
- **Mapeamento dos descritores de consulta de uma base de dados para outra** (comutação);
- Prover uma **linguagem de busca comum** que sirva de mapa para múltiplas bases de dados;
- Prover uma **linguagem de indexação comum** para um número de bases de dados em um campo;
- **Mapeamento de descritores de indexação de uma base de dados para outra.**

- **Suportar processamentos de documento após a recuperação**
  - Por exemplo: Destacar descritores responsáveis pela recuperação, usando diferentes cores para diferentes facetas.
  - **Destacar** termos pertencentes a uma dada categoria, por exemplo, **nomes pessoais**, também usando cores para diferentes categorias.
  - **Preparar sumários de documentos**, possivelmente em diferentes línguas, levando em conta os tópicos de consulta.
  - **Tradução de documentos completos.**

- **Extrair facetas dos textos.** Compilar e organizar facetas extraídas de diversos textos.

- **A função básica de base de conhecimento em conceitos e terminologia.**

  - **Mapear o espaço dos conceitos, relacionar conceitos para termos, e prover definições,** deste modo provendo orientação e servindo como uma ferramenta de referência.

  - Prover um **mapa semântico e uma linguagem comum** para um campo individual e, talvez mais importante, mapear os relacionamentos entre campos.

  - **Clarificar conceitos ao colocá-los em um contexto de uma classificação** / tipologia e para prover um sistema de definições.

  - **Relacionar conceitos e termos entre disciplinas, linguagens e culturas.**

3.2.2 Técnicas para utilização dos tesauros na Web

As funções listadas no capitulo anterior são funções potenciais. A utilização de tesauros para exercer essas funções exige a adaptação dos sistemas à estrutura do tesauro. Além disso, algumas técnicas foram desenvolvidas para exercer determinadas funções e são úteis nessa adaptação. As técnicas encontradas na literatura foram:

- Indexação
- Indexação automática
- Técnicas para melhora dos resultados de busca
- Navegação
- Técnicas específicas para bases de dados

3.2.2.1 Indexação

Os tesauros são utilizados tanto para a indexação humana ou intelectual como na indexação automática na Web.

Os objetivos dos tesauros para Indexação humana ou intelectual na WWW são relacionados por Naumis (2001) da seguinte maneira:

1. Servir de vocabulário oficial para coordenar dois processos: a indexação e a recuperação dos documentos digitais de um sistema.

2. Propor um sistema de símbolos lingüísticos para agrupar informação similar relacionada ou guiá-la para grupos mais específicos ou mais gerais de uma temática.

3. Obter uma normalização da terminologia do sistema de informação em que será utilizado.

4. Propor um conjunto estruturado de termos sobre a base de um sistema de conceitos aptos para organizar os conteúdos dos sistemas.

Uma das formas de aplicação prática do uso de tesauros na Indexação de recursos Web é a utilização dos metadados desses recursos. Os metadados, de forma mais simples, são dados sobre os dados. Para Milstead (1998), os esforços feitos para o desenvolvimento de metadados teriam um significativo impacto nos tesauros. Os formatos de metadados são interessantes por prover um modo para especificar a autoridade usada para o conteúdo da marcação e com isso auxiliar o acesso ao recurso pelo usuário. Além disso, produtores interessados em prover acesso por assunto para seus recursos usariam tesauros para a determinar o conteúdo utilizado no metadado de assunto.

Segundo a revisão ANSI/NISO Z39.19-200X, metadados podem ser usados com vocabulários controlados em diversos modos:

- Usando um vocabulário controlado como uma fonte para termos permitidos para um elemento de metadados em particular. Muitos conjuntos de metadados existentes suportam metadados relacionados a assunto como um campo “palavra-chave” ou “assunto”. Freqüentemente o conjunto de elementos, ou uma implementação comunitária particular de um conjunto de elementos, deverá indicar se um vocabulário controlado pode ou deve ser usado para um elemento particular de metadados. Nesse caso, o vocabulário controlado é usado para selecionar metadados descritivos sobre o recurso de conteúdo.

- Usar metadados para descrever um vocabulário controlado como um todo para descoberta de recursos. Esse uso de metadados não é diferente que descrever qualquer outro tipo de recurso. Muitos conjuntos e esquemas de metadados existentes podem ser utilizados para descrever um recurso de vocabulário controlado.

- Usar metadados e esquemas de metadados para representar o conteúdo integral do vocabulário controlado. Esse uso de metadados é geralmente projetado para facilitar a busca ou exportação do vocabulário controlado. Isso necessita de um mínimo: um conjunto de elementos de metadados para descrever os conceitos, termos e relacionamentos; um conjunto de definições; e um esquema de metadados para representar relacionamentos entre termos.

O interessante do item 2 é que pode ser feita a indexação do próprio tesauro, pois ele também é um recurso presente na Web. E o item três abrange o uso de XML e RDF, pois os metadados são as marcações usadas nessas linguagens para representação das relações.

3.2.2.2 Indexação automática

Os tesauros podem ser usados para a indexação automática de recursos Web. Para isso, podem ser utilizados tanto os tesauros construídos automaticamente como os tesauros construídos por especialistas, mas representados em um formato legível por computadores como o XML/RDF.

E importante ressaltar que os resultados obtidos somente com a indexação automática apresentam limitações em relação à qualidade das indexações por causa de características da linguagem natural e de falta de possibilidade de interpretações, além do fato de se basear em documentos desestruturados semanticamente.

3.2.2.3 Técnicas para melhora dos resultados de busca

3.2.2.3.1 Expansão de consulta

Salton & McGill (1983 citados por Mandala, Tokunaga & Tanaka, 2000) afirmam que um dos maiores problemas na recuperação da informação é a dificuldade de descrever as necessidades do usuário em termos de uma consulta, de modo que o sistema possa precisamente distinguir entre documentos relevantes e irrelevantes. Como conseqüência

disso, a consulta original declarada pelo usuário irá geralmente consistir de alguns poucos termos relacionados ao assunto de interesse[18]. Resumidamente, a má formulação de consulta não traz bons resultados na pesquisa. Para a resolução desse problema, Ekmekcioglu, (1992) e Fox, (1980) citados por Mandala, Tokunaga & Tanaka (2000) afirmam que ***Query expansion***[19] é técnica mais apropriada a ser utilizada. A expansão da consulta é feita pela adição de termos que são proximamente relacionados com o termo original de consulta (Mandala, Tokunaga & Tanaka, 2000). Esses termos de expansão podem ser selecionados por meio de referência a tesauro (Crouch, 1990; Paice, 1991; Crouch & Yang, 1992; Jing & Croft, 1994; Kristensen, 1993; apud Mandala, Tokunaga & Tanaka, 2000 & Milstead, 1998, Shiri & Revie, 2001, e Hudon, 2003) ou por meio de consultas aos usuários usando técnica de retorno de relevância[20] (Salton & Buckley, 1990; Buckley & Salton, 1994 apud Mandala, Tokunaga & Tanaka, 2000). Pesquisas passadas verificaram a efetividade do retorno de relevância, mas isso coloca a obrigação no usuário para certas extensões. Além disso, se o usuário não estiver familiarizado com o vocabulário da coleção de documento, será difícil ele obter bons termos de expansão, a menos que o sistema possa sugerir termos ao usuário.

Milstead (1998) complementa afirmando:

> "Uma óbvia forma em que o tesauro pode ser aplicado diretamente na recuperação é usá-lo como uma forma de expandir a busca. Pesquisas, todavia, mostram que esses relacionamentos precisam ser usados com cuidado. Em geral, expandir uma busca para incluir os termos específicos tende a melhorar a revocação sem grande sacrifício na precisão. Expandir para incluir termos mais gerais ou relacionados, embora melhore a revocação, tipicamente tem um impacto negativo na precisão." (MILSTEAD, 1998)

Para a *Query Expansion*, são usados 2 tipos de tesauros:

- Tesauros construídos por especialistas[21], e (FOX, 1980 apud MANDALA, TOKUNAGA & TANAKA, 2000)

---

[18] Esse problema também é citado por Marchionini (1989 apud Shiri, Revie & Chowdhury, 2002, p.113).

[19] Q*uery expansion* pode ser traduzida por expansão da consulta, porém não acreditamos que o termo escolhido consiga exprimir o real significado do termo em inglês.

[20] Tradução nossa a partir do termo "*Relevance feedback technique*".

[21] Tradução nossa a partir do termo "*Hand-crafted thesauri*".

- Tesauros construídos automaticamente (CHEN ET AL, 1995; CROUCH, 1990; CROUCH & YANG, 1992 apud MANDALA, TOKUNAGA & TANAKA, 2000)

3.2.2.3.2 Rankeamento dos resultados de busca

Silveira e Ribeiro Neto (2004) utilizaram os conceitos presentes nos tesauros para melhorar os resultados de busca. Para isso, os termos usados na consulta são usados para coincidir com os conceitos no tesauro e esses conceitos são usados para encontrar outros conceitos relacionados que são interpretados como fontes independentes de conhecimento evidencial. Cada fonte de evidência é usada para produzir um ranking separado baseado em conceito dos documentos nessa coleção. Esse ranking parcial será combinado em um ranking final. Desta forma, o tesauro serve para rankear os resultados da busca.

3.2.2.4 Navegação

Uma das técnicas mais utilizadas para o uso de tesauros é a construção de sistemas de navegação por conteúdo por meio da utilização dos tipos de interfaces apresentadas no capítulo 3.1.2.

A navegação permite a exploração de uma ou várias bases de dados ou a criação de mapas de assuntos de sites Web.

3.2.2.5 Técnicas específicas para bases de dados

Como praticamente todas as principais bases de dados apresentam interface Web, é pertinente citarmos a técnicas específicas para o uso de tesauros em bases de dados.

3.2.2.5.1 Descoberta de conhecimento em bases de dados bibliográficas

Pierret et al. (2005) utilizaram o tesauro MESH para descoberta de conhecimento em bases de dados bibliográficas ou (em francês: *Découverte de Connaissances dans les Bases de Données Bibliographiques* ou em inglês: *Knowledge Discovery in Databases – KDD*). Esse método utiliza o tesauro para comparar palavras-chave e com isso otimizar a recuperação de documentos pertinentes. São considerados pertinentes os documentos que utilizem as mesmas substâncias ou os mesmos sintomas ou as mesmas doenças. O tesauro serve para tratar a informação antecipadamente, evitando um maior trabalho do pesquisador. Em cima

dos documentos recuperados, utilizam o método de comparação de Swanson, com objetivo de criar comparações entre causas, doenças e medicações.

3.2.2.5.2 Tesauros melhorando interfaces de busca de bases de dados

Os tesauros ajudam aos usuários finais através do design de sistemas usando tesauros conforme afirmam Shiri, Revie & Chowdhury (2002)

> "Tradicionalmente, tesauros são usados por especialistas em busca para selecionar termos de busca alternativos para melhorar os resultados. Recentes desenvolvimentos em busca pelo usuário final e a enorme disponibilidade de sistemas de recuperação da informação online juntamente com o design de interface centrado no usuário tem aberto novos horizontes para utilização de tesauros como ajuda na busca para usuários finais" (SHIRI, REVIE & CHOWDHURY, 2002, p.11)

Eles afirmam que a importância da interface como suporte na busca da informação em geral e na seleção de termos em particular tem dado ênfase os modelos de interação em recuperação da informação. No centro de todos esses modelos encontra-se o processo de *query formulation*[22]. Como em ênfase dada por Saracevic (apud Shiri, Revie & Chowdhury, 2002), a seleção dos termos de busca para *query formulation* é dinâmica, o processo interativo necessita de uma grande variedade de facilidades e características de interface para como suporte aos usuários, de modo a facilitar o processo.

O trabalho de Shiri, Revie & Chowdhury (2002) teve como objetivo fazer revisão da literatura de modo a cobrir os esforços para integrar o padrão tesauro como parte de interfaces de busca de sistemas de recuperação da informação que objetivavam ajudar os usuários na seleção de termos de busca para *query formulation e expansion*. As facilidades mais promissoras para promover melhoras no processo de busca para o usuário final foram sumarizadas na lista abaixo:

- Uma explicita opção de busca no tesauro na principal página de busca é um caminho fácil de uso para usuários finais. Termos como "termos sugeridos", "tesauro" e

[22] É possível ser traduzido por Formulação de consulta.

“cabeçalhos de assunto” devem ser usado para mostrar a disponibilidade da facilidade do tesauro na interface.

- Fornecer uma terminologia fácil e compreensível para descrever os relacionamentos entre descritores e termos. Em algumas interfaces os relacionamentos entre termos são mostrados usando anotações como NT, BT, RT, USE etc. Outros têm usado a forma completa dos relacionamentos dos tesauros como termo geral, específico e termos relacionados. Existem também algumas interfaces que tem usado sinais como “+” e “-“ para demonstrar os relacionamentos genéricos e específicos e termos.

- Fornecer listas alfabéticas, hierárquicas e permutadas para suportar diferentes estratégias de navegação e busca.

- Modos flexíveis para escolha de termos para postar para o sistema de busca como “arrastar e colar”, caixas de seleção, características de hipertexto e duplo clique.

- Facilitar o processo e entendimento da movimentação de um descritor para sua estrutura hierárquica usando navegação em hipertexto.

- Fornecimento para a seleção de operadores Booleanos alternativos para combinação de diferentes termos do tesauro.

- Prover um retorno dos termos não disponíveis em um tesauro e sugerir termos relacionados em certa quantidade para o termo inicial consultado.

- Prover uma opção *“term pool”* para salvar os descritores escolhidos pelos usuários durante a navegação do tesauro para uso posterior.

- Integrar apresentação de documentos recuperados e tesauro para uma busca e recuperação mais efetiva.

- Disponibilidade da opção do tesauro em todos os estágios do processo de busca, a saber formulação de busca, modificação ou expansão.

3.2.3 Ambientes informacionais em que os tesauros são utilizados

A Web é uma reunião de diversos tipos ambientes informacionais, com distintos objetivos. Nosso objetivo foi descobrir em que tipos de ambientes os tesauros estão sendo empregados. Encontramos na literatura os seguintes tipos de sistemas:

3.2.3.1 Bases de dados ou Sistemas de Recuperação da Informação

Para Hudon (2003), os tesauros que estão completamente integrados em bases de dados têm o uso mais imediato para a recuperação da informação na Web, pois tesauros que operam em conexão com uma base de dados oferecem suporte avançado para os usuários ou buscadores de informação. Isso acontece porque o tesauro foi usado na indexação do conteúdo da base de dados por profissionais indexadores na maioria dos casos. Isso garante a consistência e melhora a precisão da recuperação da informação para o usuário final.

Os tesauros integrados em SRI são o uso mais freqüente de tesauros na Web e é semelhante ao uso feito em SRI tradicionais. Eles normalmente são utilizados com técnicas de indexação, navegação, *query expansion*, entre outras.

Além disso, a literatura aponta sistemas semelhantes:

3.2.3.1.1 Sistemas de busca e navegação multi-tesauros

Shiri & Revie (2001) afirmam que os sistemas de busca e navegação multi-tesauros aparecem a partir da comprovação que o uso de vocabulários controlados melhora a qualidade e disponibilidade como suporte para buscas em várias bases de dados e de que isso acontece também quando nos movemos para uso de diferentes tesauros para busca em base de dados cruzadas, com isso, os sistemas de busca e navegação multi-tesauros usam diversos tesauros para a busca e navegação em bases de dados.

Porém para que possa ser usado amplamente, é necessário resolver o problema da falta de interoperabilidade:

> “Esse entusiasmo para usar a moderna tecnologia da Web para publicar tesauros na Web resultou em um crescente número de tesauros e a necessidade para pensar em interoperabilidade de tesauros como uma

> necessária para acessar e usar diferentes tesauros para busca e recuperação" (Shiri & Revie, 2001)

Um exemplo de multi tesauros é o *Unified Medical Language System (UMLS) Metathesaurus.* O multi tesauro UMLS é usado em uma grande variedade de aplicações incluindo: linkagem entre diferentes vocabulários clínicos ou biomédicos; recuperação da informação de bases de dados com termos de cabeçalhos de assuntos especificados por humanos e fontes de informação em texto livre; linkagem de registros de pacientes a informações relacionadas na bibliografia; texto completo ou bases de dados efetivas; processamento de linguagem natural e pesquisa em indexação automática e entrada de dados estruturados.

Há também os sistemas de gerenciamento de multi tesauros com interface Web. Shiri & Revie (2001) afirmam que Sistemas de gerenciamento multi tesauros com interface web são também outro novo desenvolvimento usando múltiplos tesauros. O objetivo do projeto por eles analisado é prover um modo para buscar em bases de dados distribuídas de medicina alternativa produzidas em vários países. O sistema de gerenciamento do tesauro possui dois níveis, ambos com uma interface Web: uma busca do site aberta a qualquer pessoa que queira buscar ou navegar o tesauro cruzado e um site de manutenção do tesauro para a sua edição.

3.2.3.2 Subject-based information gateways

Conforme Shiri & Revie (2001), os tesauros podem ser empregados em "*subject-based information gateways*". Koch (2000 apud Shiri & Revie, 2001) define subject gateways como:

> *"Internet-based services which support systematic resource discovery. They provide links to resources (documents, objects or services), predominantly accessible via the Internet. Browsing access to the resources via a subject structure is an important feature."*[23] (KOCH, 2000)

[23] Tradução nossa: "Serviços baseados na Internet que suportam descoberta sistemática de recursos. Eles provêm links para recursos (documentos, objetos ou serviços), predominantemente acessíveis via Internet. O acesso por navegação aos recursos por meio de estrutura de assunto é uma importante característica."

Para eles, o acesso por assunto em alguns tipos de estrutura de conhecimento como tesauros e sistemas de classificação é uma das mais significantes características de uma boa *subject gateway*. Esse controle de qualidade de *subject gateways* tem estabelecido procedimentos para seleção e descrição de conteúdo de páginas web e também uso de tesauro para cuidadosa e consistente descrição de conteúdo. Recentemente, diversos *subject-based information gateways* têm sido desenvolvidos na Web com o uso de tesauros para indexação e recuperação de paginas e Web sites. Seguem alguns exemplos:

- *Art, Design, Architecture and Media information gateway (Art and Architecture thesaurus);*
- *Engineering Electronic Library, Sweden (Engineering Information's EI thesaurus);*
- *Organising Medical Networked Information (Medical Subject Headings (MeSH) thesaurus);*
- *Social Science Information Gateway (HASSET thesaurus).*

Esses *subject gateways* usam tesauros para indexar páginas Web e prover acesso por assunto mais consistente e estruturado para navegação e busca de páginas Web.

3.2.3.3 Bibliotecas Digitais

Os tesauros estão sendo usados em Bibliotecas Digitais, conforme Hodge (2000) afirma:

> "Sistemas de organização do conhecimento (KOS) podem melhorar a biblioteca digital de diversos modos. Eles podem ser usados para conectar um recurso da biblioteca digital a um recurso relacionado. A informação relacionada pode residir no próprio KOS ou o KOS pode ser usado como um arquivo intermediário para recuperar a chave necessária para acessar ele em outro recurso. Um KOS pode tornar materiais digitais acessíveis para comunidades diferentes. Isso pode ser feito através do provimento de um alternativo acesso por assunto, por adicionar acesso por diferentes modos, provendo acesso multilíngüe, e usando o KOS para suportar buscas em texto completo." (HODGE, 2000)

Soergel (2002) aponta os seguintes itens que as bibliotecas digitais podem melhorar com o uso de tesauros:

- Melhorar recuperação efetiva para manipular a crescente massa de materiais.
- Prover acesso unificado aos materiais em diferentes mídias (especialmente acesso a materiais não textuais)
- Prover suporte de conhecimento para usuários finais que acessam informação em rede sem o benefício de um intermediário.
- Suportar a criação e manutenção de sistemas de informação personalizados ou de grupos de trabalho.
- Suportar busca pela informação como uma parte integral de solução de problemas, aprendizado e trabalho intelectual.
- Suportar trabalho colaborativo.
- Suportar busca da informação como uma parte integral para a solução de problemas, aprendizado e trabalho intelectual.
- Ajudar usuários a explorar idéias em conjunção com a exploração da informação.
- Suportar recuperação fina e assimilarização da informação.
- Suportar processamento da informação junto com ou após a recuperação.

3.2.3.4 Blogs

Gammel (2005) descreve algumas possibilidades de emprego de tesauros em Blogs:

- Blogueiros da Internet usam termos do tesauro para criar categorias para seus blogs. Leitores de uma Internet, por exemplo, pode então ver posts de blogs criados por qualquer um na rede para um termo particular do tesauro. Links para categorias relacionadas, gerais e específicas podem ser criados automaticamente. Essencialmente um meta blog de conteúdo baseado em termos do tesauros mais usados freqüentemente.

- A idéia precedente pode também ser feita através da determinação de termos de tesauros para entradas individuais de blogs e então indexar esses metadados.

- Um índice hierárquico de assunto de blogs pode ser criado baseado nas categorias que são usadas por escritores individuais de blogs. Eles incluirão mais categorias quanto escreverão conteúdo nessas áreas.

- Um diretório/índice como o Yahoo!© de uma intranet pode ser criado baseado no tesauro que indexa um conjunto de conteúdos blogados. O efeito "bombardeando o Google©" dos blogs então aumenta mais conteúdos relevantes nos primeiros da lista de resultados da busca.

- Blogs indexados por uma estrutura de tesauro tornam muito mais fácil encontrar outros blogs sobre tópicos similares sem ter que confiar nos próprios blogueiros para criar associação via links diretos. Isso pode ser uma ferramenta suplementar para referencias que correntemente direciona tráfico entre blogs.

- O gerente de tesauro pode monitorar blogs relacionados para nova linguagem ser usada isso pode ser adicionado em um tesauro como um termo formal.

3.2.3.5 Web Sites

Rosenfeld & Morville (2001) utilizam os relacionamentos presentes nos padrões de tesauros para a construção de sistemas de navegação em Web Sites pela área da "Arquitetura da Informação":

> *"Metadata and controlled vocabularies present a fascinating lens through which to view the network of relationships between systems. In many large metadata-driven web sites, controlled vocabularies have become the glue that holds the systems together. A thesaurus on back end can enable a more seamless and satisfying user experience on the front end"*[24] *(ROSENFELD & MORVILE, 2001)*

[24] Tradução nossa: "Metadados e vocabulários controlados apresentam uma fascinante lente para ver a rede de relacionamentos entre sistemas. Em muitos amplos web sites que são dirigidos por metadados, vocabulários controlados se tornam a cola que mantém o sistema junto. Um tesauro no *'back end'* pode permitir uma maior experiência do usuário final no *'front end'*."

Porém ele indica que os tesauros foram desenvolvidos para Bibliotecas, museus e agencias governamentais antes da criação da World Wide Web e por isso não é possível ser copiado indiscriminadamente pelos Arquitetos da Informação.

Eles relatam que atualmente poucos times de arquitetos da informação possuem conhecimento ou suporte para esse significante investimento, mas espera que isso mude em poucos anos: “o tesauro se tornará uma ferramenta chave para administração com o crescente tamanho e importância dos web sites e intranets.”

O trabalho de Rosenfeld & Morville (2001) descreve os tesauros e os exemplos de utilização de tesauros na Web, mas como para o seu uso é necessária a adaptação às necessidades do web site ou intranet, o uso do tesauros ainda é feito de forma empírica.

Já Hassan Montero & Núñes Peña (2005) apresentam um modelo mais prático para o emprego de tesauros na Arquitetura das informações de Web sites, conforme observamos na figura abaixo:

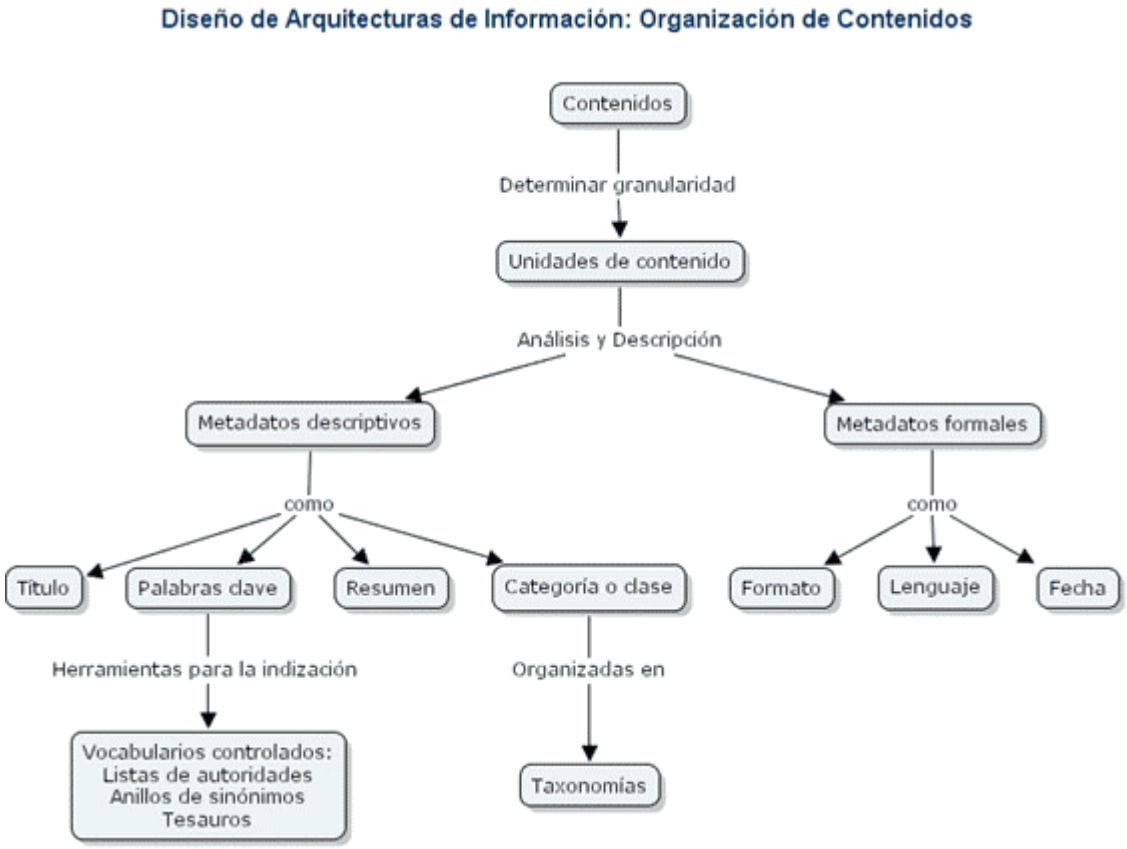


Figura 1 : Desenho de Arquiteturas de informação: Organização de Conteúdos

Fonte: HASSAN MONTERO & NÚÑES PENA (2005)

Os tesauros e os vocabulários controlados são utilizados para a indexação intelectual ou humana de palavras-chave nos metadados descritivos para evitar a sinonímia e a polissemia e também serve como opção de navegação para o usuário.

3.2.3.6 Intranet corporativa

As Intranet apresentam características tanto de sistemas de recuperação da informação quanto da Internet. Méndez Rodriguez (2000) descreve a importância de uso de tesauros para a resolução do problema de recuperação da informação em texto completo no ambiente web corporativo:

- As Intranets se desenvolvem segundo os mesmos padrões que a Internet (HTML, XML, etc.) e como a Internet, é normalmente um conjunto de recursos descentralizados. Contudo, as Intranets supõem limites finitos – ou ao menos previsíveis – de informação, além de ter uma maior homogeneidade temática e uma complexidade de tipos de informações controláveis. Essas características fazem que a Intranet possa assumir com mais facilidade o objetivo da organização e recuperação da informação.

- Por outra parte, os sistemas de recuperação de informação na Internet de propósito geral (Altavista, Northenlight, etc) se baseiam na extração automática da informação e carecem de técnicas de gestão do conhecimento e portanto não podem dar uma resposta precisa a uma pergunta concreta sobre o conteúdo semântico dos documentos, e por isso recuperam tanto ruídos. Contudo, todos os sistemas de recuperação de informação de qualidade na rede – os denominados *subject gateways*, que prefiro chamar de “sistemas de recuperação de informação de organização bibliotecária” – que centralizam seus esforços na seleção, descrição e organização de recursos de uma área temática. Somente em contextos muito concretos de recuperação de informação na Internet se utilizam normas de valor semântico como vocabulários ou tesauros para descrever o conteúdo dos documentos como para realizar as buscas.

- Enquanto que a Internet é um ambiente infinito, multilíngüe e heterogêneo, uma Intranet é em si mesma um sistema de informação temático, uma *subject gateway* de visibilidade limitada, finita, mais homogênea e tipificável e normalmente mono/bilíngüe. Por isso parece ser um ambiente informativo apropriado para basear a recuperação da informação em sistemas de organização do conhecimento como tesauros e classificações, que normalizem os atributos dos metadados descritivos aplicáveis.

E com isso, ela defende o uso de Metadados e vocabulários controlados para a criação de sistemas de recuperação da informação na Intranet similares aos *subject gateways* da Internet.

3.3 Problemas encontrados para a utilização de tesauros

Durante o trabalho, encontramos na literatura diversos problemas que ainda impedem uma maior utilização dos tesauros:

Shiri & Revie (2001) indicam uma falta de padrões para a publicação de tesauros na Web que causa problemas em relação à interoperabilidade, reusabilidade e compartilhamento de tesauros e afirmam que existe uma necessidade urgente para examinar as ferramentas semânticas e sintáticas, formatos e padrões, usados por editores de tesauros baseados na Web e para buscar meios em que esses aspectos possam ser harmonizados ou integrados. Além disso, eles afirmam que muitos tesauros baseados na Web não são completamente integrados como ajuda na busca e navegação em base de dados, sistemas de recuperação da informação e ferramentas de busca na Web emergentes. Essas ferramentas podem ser efetivamente utilizadas pelas máquinas de busca para mais consistentes e unificadas descrições de recursos e descobertas e ainda os tesauros baseados na Web podem também ser considerados ferramentas para formulação de consultas, refinamento e expansão e ajudar usuários a definir mais precisamente e claramente as necessidades de informação. Porém, esforços são requeridos para estimar a extensão que essas ferramentas poderão contribuir para recuperação mais efetiva e confiável no contexto da Web.

Além disso, os tesauros atuais possuem diversas limitações que podem impedir um emprego mais efetivo na automatização de sistemas de informação. Soergel et. Al (2004) sumarizou as limitações dos atuais tesauros da seguinte maneira (grifos do autor):

- **Falta de uma abstração conceitual**: tesauros e outros KOSs tradicionais são coleções de termos (genéricos ou de um domínio específico), organizados em uma estrutura poli hierárquica ou uma estrutura arbórea mono hierárquica e interligada com alguns relacionamentos muito gerais e básicos. A distinção entre um conceito (significado) e sua lexicalização (palavras) não cria consistência, se em tudo, em um sistema, e como tal ele não reflete o modo humano de entender o mundo em termos de significado e linguagem.

- **Cobertura semântica limitada**: a maioria dos tesauros não diferencia conceitos em tipos e têm um conjunto muito limitado de relacionamentos entre conceitos, distinguidos somente entre relacionamentos hierárquicos e relacionamentos associativos. Esses relacionamentos muito rudimentares não têm poderes para guiar o usuário na descoberta de informação através de significados na Web ou suportar inferência. Eles não refletem os relacionamentos conceituais que as pessoas conhecem e que podem ser usados por um sistema para sugerir conceitos para expandir a consulta ou torná-la mais específica.

  As relações entre conceitos providas pela maioria dos tesauros força todas as relações em duas categorias gerais: hierárquica e associativa. Muito freqüentemente os relacionamentos semânticos capturados deste modo são ambíguos e pobremente definidos. A generalização/especialização das relações definidas em muitos tesauros não são adequadamente desenvolvidas para serem usadas para descrição semântica e descoberta de recursos Web. Então existe a necessidade para um rico e mais poderoso conjunto de relacionamentos.

- **Falta de consistência**: devido à falta de precisão semântica dos relacionamentos nos tesauros, eles são aplicados inconsistentemente, criando ambigüidade na interpretação dos relacionamentos e resultando em uma estrutura semântica interna total que é irregular e não prognosticável. Muitas dos relacionamentos hierárquicos NT/BT podem, por exemplo, serem resolvidos para relacionamentos RT não hierárquicos, e vice versa.

- **Limitado processamento automático**: tradicionalmente tesauros são projetados para indexar e formular consultas por pessoas e não para processamento automatizado. A

semântica ambígua que caracteriza muitos tesauros os torna não adequados para processamento automático.

Esses problemas levam a necessidade de um novo padrão de tesauros, conforme afirma Hudon (2003):

> "A new standard should be submitted to the community as quickly as possible, however, if the goals of conceptual and technological compatibility are to be kept within reach."[25] (HUDON, 2003, p.118)

E complementa:

> "A Segunda geração de tesauros é agora, realmente, necessária. O novo tesauro deverá ser desenvolvido como a necessidade dos usuários e hábitos em mente e ser estruturado para que ele possa ser usado mais eficientemente em ambientes informacionais guiados por ferramentas de busca. Muito tempo é atualmente desperdiçado tentando convencer varias categorias de gerentes Web e usuários que uma levemente modificada versão do tradicional tesauro, uma ferramenta que permanece muito cara para desenvolver e preservar, é alguma coisa que eles necessitam absolutamente. Igualmente mais tempo deve ser devotado para o design e teste de modelos realmente novos, ricos, e mais versáteis. No nível semântico, o tesauro do futuro precisa oferecer relacionamentos mais definidos, especificando a natureza das ligações entre termos. No nível prático, a segunda geração de tesauro pode ser projetado e oferecido em uma forma mais interativa, sobre as quais Bertrand-Gastaldy e Davidson sugerem que qualquer um pode prognosticar que tesauro precisa eventualmente ser usado em redes globais, por não especialistas e para outras propostas que somente indexação e recuperação da informação." (HUDON, 2003, p.118)

E Sajus (2002) complementa, afirmando a necessidade de se transformar o tesauro em uma ferramenta automatizada:

---

[25] Tradução nossa: "Um novo padrão deve ser enviado o mais rápido possível para a comunidade, porém somente, se os objetivos da compatibilidade conceitual e tecnológica tiverem sido alcançadas"

> "A função tesaural que se sustenta de dezenas de experiências sobre o difícil terreno do acesso de informação por questão, constitui uma fonte preciosa para invenção da Web Semântica. É portanto urgente à necessidade de renovar as normas tesaurais, reposicionando-as em relação às novas ferramentas e métodos de gestão semântica. Essa atualização deverá orientar o tesauro não mais na direção da indexação manual e sim na direção de tratamentos automáticos e semi-automáticos da linguagem toda ao desenvolver sua função heurística. Dentro dessa perspectiva que é se faz necessário visualizar o futuro da função tesaural no coração dos sistemas de informação" (SAJUS, 2002)

Shiri & Revie (2001) citam um workshop[26] que teve como objetivo investigar o desejo e a possibilidade de um padrão para tesauro eletrônico e que chegaram aos seguintes tópicos:

- Fale sobre critérios e/ou métodos para geração de tesauros por meio de ajuda de máquinas ou meios automáticos;
- Mostrar relações semânticas entre termos, como ajuda para texto e análise e recuperação da informação;
- Suportar uma variedade de apresentações do tesauro eletrônico;
- Suportar protocolos de interoperabilidade, estruturas, e/ou semânticas aplicáveis aos tesauros.

Já Soergel et al. (2004), propõem que para superar as limitações e criar capacidade para uma busca mais poderosa e processamento inteligente de informação, especialmente com essas capacidades serem mais amplamente disponíveis na Web, KOSs tradicionais precisam ser re-projetados em KOS que contém ligados conceitos de domínios em um rico *network* de relacionamentos bem definidos e um rico conjunto de termos identificando esses conceitos:

> "Em contraste aos tradicionais KOS, ontologias provêm abstração conceitual e relacionamentos diferenciados. Ontologias especificamente separam

[26] "Electronic Thesauri: Planning for a Standard" promovido por National Information Standards Organization (NISO), American Society of Indexers (ASI) e Association for Library Collections and Technical Service (ALCTS) em novembro de 1999.

> conceitos de lexicalizações e isso reflete melhor a estrutura do entendimento humano de um domínio. Em ontologias, as semânticas são desenvolvidas por meio de assegurar que cada conceito em um domínio é único e precisamente definido e por especificar relacionamentos elaborados entre os conceitos. Esses relacionamentos em uma ontologia são explicitamente nomeados e desenvolvidos com especificação de regras e obrigam deste modo que ele reflita o contexto de um domínio para qual o conhecimento é modelado. Dando essa semântica mais precisa e sem ambigüidade, ontologias permitem favorecer a inferência do conhecimento pela representação explicita do conhecimento na ontologia. O novo (implícito) conhecimento pode ser derivado na aplicação de generalização ou regras transitivas, o nível da aplicabilidade que é limitado em um pobremente definido KOS como um tesauro. Esse conhecimento adicionado na ontologia torna possível isso quanto empregado para processamento inteligente de informação. Contudo existe um alto custo envolvido na transformação de tesauros para ontologias, existe uma expectativa que adicionar poder de consistência, precisão, e torná-lo completo será merecedor do investimento ainda que um grande número de retorno de investimento do desenvolvimento de ontologias é difícil de retornar.“ (SOERGEL ET AL, 2004)

Pincemin complementa (2003) com a comparação entre ontologias e tesauros e conclui que ontologias são para representação e tesauros documentários são para mediação no contexto da Web Semântica.

Essas necessidades apontadas na literatura não foram completamente supridas na revisão da norma ANSI/NISO Z39.19-200X, principalmente no que se refere à automatização.

A literatura analisada aponta diversos benefícios potenciais de uma futura geração de tesauros. Esses benefícios partem sempre do princípio da necessidade de automatização, alguns com visão mais radical, como Soergel, outras mais amenas, como Miles. Os benefícios apontados por eles foram listados da seguinte maneira:

Soergel et al. (2004) afirma a necessidade de melhorar a organização e recuperação da informação de modo que satisfaça os usuários, uma vez que não é possível pelo uso dos tradicionais KOS e aponta os benefícios potenciais dos KOS caso as mudanças propostas por ele, isto é, a reengenharia do tesauro de modo que se aproxime da concepção de ontologia, sejam feitas:

- **Identificadores únicos e semântica formal:** a explicita definição de conceitos e relações em uma ontologia permite um único identificador ser designado para cada conceito. Como cada conceito e relação é explicitamente definida como uma única entrada, a ontologia serve para a formalização semântica.
- **Consistência interna:** outro beneficio de semânticas explicitas é a realização de uma consistência estrutural interna na expressão do conhecimento dada à possibilidade de aplicar limites integralmente.
- **Interoperabilidade:** semânticas claras permitem interoperabilidade entre diferentes KOSs desde conceitos correspondentes em diferentes KOSs possam ter o mesmo único identificador, sem restrição das lexicalizações atuais usadas para expressar esses conceitos. Interoperabilidade semântica promove troca e reuso do conhecimento.
- **Ótima integração da informação**: interoperabilidade entre diferentes KOSs torna isso possível para máquinas reconhecer e analisar o pretendido significado dos termos de vocabulários diferentes. Isso é possível pelo uso de meta informação estruturada e descrição formal do conhecimento como esquemas agregados de metadados, vocabulários de domínios controlados e taxonomias. A capacidade para integrar terminologias de diferentes fontes maximiza o valor do investimento feito na ontologia.
- **Capacidade de inferência**: novos KOSs têm o potencial para expressar conhecimento em torno o que é apresentado na estrutura do sistema. Ao contrário dos tradicionais KOS onde conceitos e relações são pouco ou não especificadas, se qualquer regra axiomática existir, os fatos (conceitos e relações) e regras que podem ser derivadas de uma ontologia têm as capacidades expressivas que permitem o raciocínio.
- **Processamento automatizado da informação**: novos KOSs criam um potencial de melhora para descobrir informação relevante de diferentes fontes pela exploração de modelos e filtragem de informação usando conexões conceituais representadas na ontologia. Isso possibilita a respostas a questões em uma ou mais bases de dados, ou usar processamento de linguagem natural, pelo texto.
- **Suporte para o Processamento de linguagem natural (NLP)**: oferece a possibilidade de provir uma resposta direta para uma questão de busca expressada em linguagem natural, usando os relacionamentos e semânticas melhoradas em uma ontologia, em vez de somente retornar uma lista de documentos relevantes.

- **Entendimento das consultas de busca**: usando NLP e processamento semântico, um sistema pode entender uma questão colocada em linguagem natural, determinar os conceitos envolvidos e, quando necessário, criar consulta Booleana.
- **Busca baseada em conceitos:** uma ontologia pode prover capacidades especificas de busca com conhecimento do contexto para uma área de interesse.
- Suporte para busca/navegação de informação integrada: mineração de texto na Web (*Web mining*) através do acesso orientado a significado, dinâmica organização da informação com a possibilidade para links entre domínios.
- **Busca por expansão da consulta**: a melhora, extensão e desambiguação dos termos de consulta do usuário se tornam possível com a adição de enriquecimento do domínio e informação específica do contexto.

Enquanto Soergel apresenta os benefícios dos tesauros transformados em ontologias, Miles & Rogers (2004) listam os potenciais benefícios dos KOS atuais no formato RDF[27] da seguinte maneira:

- **Indexando um recurso Web para exposição – inclusão – na "Web Semântica":** Um usuário final humano navega por apresentações hierárquicas em um tesauro particular ou algum outro KOS a fim de selecionar termos controlados que ele/ela acreditam melhor indicar o assunto (ou alguma outra propriedade) de um recurso Web. Esses termos serão então encaixados em metadados de recursos Web.

- **Indexar um recurso Web para benefício de uma comunidade de usuários específica:** Similar ao caso anterior, mas onde o usuário é um especialista no assunto. O usuário pode desejar ser oferecido termos preferidos e não preferidos de mais de um tesauro potencial – com pontos de entrada de tesauros para cada – mais a capacidade para navegar por cada tesauro a partir desse ponto. Neste sentido, o caso é também "encontrar o tesauro certo" (como sugerido por Dave Reynolds, HP). Uma versão avançada desse cenário permite que o catalogador que queira **"criar meu próprio**

[27] O formato SKOS Core Guide, que proporcionará os benefícios listados, é um formato RDF que se aproxima do formato de ontologia OWL, mas não apresenta a mesma complexidade. Ele com isso, se torna complementar ao OWL. O OWL é orientado a lógica e o SKOS é orientado a linguagem e devido a isso, possui um poder de inferência menor e maior flexibilidade quanto à semântica. (MILES ET AL, 2004).

tesauro” para um propósito específico (e talvez compartilhar ele com outros catalogadores) através de pontos de encontro em que une mais de um tesauro, e para então compor (e salvar) um sub grupo de tesauros fundidos para uso em catalogação.

- **Suporte para consulta para melhorar a busca e navegação via Ferramentas de busca da Web:** Por busca “melhorada” nós tentamos significar melhor revocação de consulta (o termo de busca do usuário final é expandido com equivalentes sinônimos ou quase sinônimos). Por navegação nos tentamos significar suporte para o usuário para restringir/refinar seus termos de busca de modo a produzir melhor precisão/relevância nos resultados da busca.

- **Suporte para consulta para melhorar a busca e a navegação via alguma ferramenta/aplicação Web:** Similar ao caso de uso anterior, mas em um contexto de comunidade especialista e via uma ferramenta que é adaptada as necessidades desta comunidade. Dependendo do sucesso dos designs de interface do usuário, e é claro das necessidades dos usuários, essa organização de melhoramento de ferramentas pode usar a forma de expansão de consulta automatizada (invisível ao usuário), ou oferecer opções de navegação ao usuário final para ajudá-lo a refinar sua estratégia de busca. Isso implicará no mapeamento entre tesauros/KOS. Nos notamos que geralmente quando nos referimos a mapeamento entre vocabulários nós podemos estar nos referindo a uma vastidão de contextos – de um contexto “mundo aberto” (ex. mapear um esquema de categorização de blogs aberto para um diretório Web aberto) para mais especializado, contexto orientado a bibliotecas (mapeamento entre coerências, KOS gerenciados).

- **Recuperação multilíngüe de imagens:** Apesar de que visualizar imagens é independente de linguagem, metadados adicionados podem ser expressos em qualquer linguagem natural, então aqui suportes para tesauro/KOS podem ser usados para traduzir consultas de imagens relacionadas do usuário final para outras linguagens.

- **Ferramenta para analisar gramaticalmente um documento e sugerir metadados:** Um serviço pode ser usado para receber um recurso completo Web (por exemplo entradas de blogs – ou itens, que são metadados sobre entradas de blogs). Esse serviço pode extrair conteúdos apropriados do documento de modo a especular (via

automatização) do que ele trata e então sugerir termos do tesauro com o qual é possível fazer a marcação dele.

- **Suporte multilíngüe para uma comunidade especializada:** Similar ao caso de recuperação de imagens mencionado aqui, mais para uma comunidade específica. Neste caso, um usuário final pode requisitar serviços de tradução para o glossário W3C. Isso representa uma necessidade para mapeamento de termos entre línguas em contextos específicos em que termos usados podem ter significados especializados não normalmente anexado no seu uso no curso do discurso em linguagem natural.

- **Suporte de mapeamento de Schema para uma comunidade especializada:** Existe a necessidade de suporte ara busca e navegação cruzada com mapeamento entre o leque de schemas usados por serviços de provedores de dados buscáveis na Web. Isso é feito de modo a tornar as expressões de consulta para os usuários finais mais confiáveis e mapeáveis entre schemas alvo. Um Portal de ferramentas de busca cruzada, por exemplo, pode fazer bom uso dessa ferramenta de suporte. Este nível de suporte aplicado ao schema “nível coleção” como oposto ao “nível item”. O ultimo é relacionado ao assunto indexado por provedores de dados no nível do registro e é relacionado mas independente área, também referida a essa lista de casos de uso.

- **Representações recuperáveis de modelos UML usadas no desenvolvimento de software.**

4 Considerações finais

As ferramentas de vocabulário podem contribuir com a organização de ambientes informacionais na Web. Porém, para que possam ser mais eficientes, as ferramentas de vocabulário precisam de estudos mais consistentes e de aperfeiçoamento nas suas regras de construção e nos formatos de representação de tesauros na tecnologia Web.

Em relação às regras de construção, a revisão da norma ANSI/NISO Z39.19-200X foi um passo importante nesse sentido. Ela permite não apenas a construção de tesauros, mas de diversos tipos de vocabulário controlado. Essa flexibilidade é importante, pois permite uma melhor adaptação da ferramenta às necessidades dos ambientes informacionais da Web. Porém, essa revisão ainda não aborda uma questão chave apontada pelos autores, que é a automatização. Não podemos entender essa falta como uma crítica, pois a automatização ainda envolve dois problemas que devem ser mais bem estudados:

1. Faltam estudos consistentes para descobrir quais funções podem ser automatizadas e como essa automatização deve ser feita.

2. Falta uma definição clara do formato mais apropriado para a representação de tesauros em tecnologia Web.

Além disso, a possibilidade de automação abre uma discussão sobre a necessidade de tornar o tesauro uma ferramenta mais específica se aproximando da idéia de ontologia, como é defendida por Soergel, ou manter a estrutura atual como defende Miles & Rogers, pois essa flexibilidade é vista por eles como uma das suas principais vantagens.

Quanto às funções, de modo geral, que os tesauros podem exercer, identificamos três categorias principais:

1. Ferramenta de linguagem para recuperação da informação:

   O tesauro é uma ferramenta que pode ser usada como uma linguagem para a melhoria de sistemas de recuperação da informação em texto completo. Isso pode ser feito com a utilização de diversas técnicas como indexação, indexação automática, ranqueamento dos resultados de busca. Porém, é importante salientar que os tesauros

apresentam sérias limitações semânticas para uma maior automatização dos processos de recuperação da informação.

2. Mediação:

   Mediação deve se configurar como função chave para o tesauro. A riqueza lingüística dos tesauros deverá ser explorada para a criação de interfaces ricas que permitirão aos usuários uma melhor navegação e aprendizado.

3. Padronização:

   Padronização de uma área será um recurso muito importante para sistemas de comunicação informais existentes hoje na Web. A padronização possibilitará a criação de ligações e permitirá um maior acúmulo de conhecimento. Além disso a principal função do tesauro, o controle de vocabulário, ganha uma maior amplitude, ao abranger vários tesauros e o tesauro passa a servir como um guia para a exploração de várias bases de dados e não só mais como controle de vocabulário. Porém ainda problemas com interoperabilidade e mapeamento entre tesauros são problemas a serem resolvidos.

Mas, para a utilização efetiva de tesauros documentários na Web, são necessárias ferramentas para tornar essa tarefa possível. Essas ferramentas são: softwares de criação e gerenciamento de tesauro, softwares de gerenciamentos múltiplos tesauros online, e integração com os softwares de gerenciamento de conteúdo existentes, softwares de indexação automática, entre outros.

A falta de um padrão estabelecido e uma melhor compreensão das funções dos tesauros na Web ainda são barreiras para os desenvolvedores de softwares. Mas, podemos citar algumas iniciativas pioneiras no desenvolvimento de softwares para tesauros na Web[28]:

- O projeto SKOS Project: é um projeto que visa desenvolver especificações e padrões para suportar o uso de sistemas de organização de conhecimento (KOS) na Web Semântica. O projeto, que pode ser acessado em: < http://www.w3.org/2004/02/skos/

[28] Devido a limitações, citamos somente os softwares livres ou iniciativas abertas.

>, contém um padrão para representação de KOS na WS, o SKOS Core, um padrão para mapeamento entre vocabulários, o SKOS Mapping e APIs (*Application Program Interfaces*) de softwares experimentais.

- Tematres: software livre voltado para a construção de tesauros e estruturas de navegação Web, que permite exportar nos formatos Dublin Core, SKOS e ZTHES. Porém, ainda está na sua versão 0.21 e apresenta muitas limitações devido a tecnologia utilizada (PHP e MySQL), mas é um programa pioneiro. Disponível em: < http://www.r020.com.ar/tematres/ >

- Uma proposta inovadora é a *"Automatización de Tesauros y su utilización en la Web Semántica"* feita por José Ramón Pérez Agüera. (PÉREZ AGÜERA, 2004). Ela é a proposta de criação de um gestor de tesauros que utiliza ambientes distribuídos mediante serviços Web baseado em RDF. É um modelo conceitual muito consistente, que deve ser explorado e desenvolvido e que pode apresentar bons resultados.

5 Referências

ANDERSON, James D. & PEREZ CARBALLO, Jose. **Information Retrieval Design**. East Brunswick (EUA): University Publishing Solutions, 2005. Disponível em: < http://www.scils.rutgers.edu/publications/ir-design/ird2005.html >. Acesso em: 09 mar. 2005.

AUSSENAC-GILLES, Nathalie & CONDAMINES, Anne. Documents électroniques et constitution de ressources terminologiques ou ontologiques. **Information-Interaction-Intelligence**, Toulouse, v. 4, n°1, 2004. Disponível em: < http://archivesic.ccsd.cnrs.fr/documents/archives0/00/00/10/16/sic_00001016_00/sic_00001016.pdf >. Acesso em: 04 abr. 2005.

BATTY, David. WWW – Wealth, Weariness or Waste: Controlled vocabulary and thesauri in support of online information acess. **D-Lib Magazine**, nov. 1998. Disponível em: < http://www.dlib.org/dlib/november98/11batty.html >. Acesso em: 20 dez 2004.

BUSHA, Charles H. & HARTER, Stephen P. **Research methods in Librarianship: techniques and interpretation**. Academic press: New York, 1980.

CINTRA, Anna Maria Marques et al. **Para entender as linguagens documentárias**. 2ª Ed. São Paulo: Polis, 2002.

CLARKE, Dave & YANCEY, Trish. **Twenty-First Century Tools for Vocabulary Management and Indexing**. In: ANNUAL MEETING of the American Society for Information Science and Technology, 2001. Disponível em: < http://www.synaptica.com/asis_2001.asp >. Acesso em: 21 dez. 2004.

CRAVEN, Tim. **Thesaurus displays on the Web**. Ontario (Canadá), 2004. Disponível em: < http://instruct.uwo.ca/gplis/677/dispthes.htm >. Acesso em: 08 dez. 2004.

CUEVA MARTÍN, Alejandro de la. Acceso y utilización de tesauros en Internet. **Revista Española de Documentación Científica**, Valência, v. 22, n.4, 1999. Disponível em: < http://161.116.140.71/pub/fburg/docs/cueva.pdf >. Acesso em: 20/12/2004.

DENCKER, Ada de Freitas Maneti; VIÁ, Sarah Chucid da. **Pesquisa empírica em ciências humanas:** com ênfase em comunicação. São Paulo: Futura, 2001.

FURGERI, Sérgio. **Ensino Didático da Linguagem**. São Paulo: Érica, 2001.

GAMMEL, David. **Thesauri and Web Logs**. Silver Spring (EUA), 2002. Disponível em: < http://www.highcontext.com/hcarchives/2002/05/28/thesauri-and-web-logs/ >. Acesso em: 10 mar. 2005.

HASSAN MONTERO, Yusef & NÚÑEZ PEÑA, Ana. **Diseño de Arquitecturas de Información: Descripción y Clasificación.** Granada, jan. 2005. Disponível em: < http://www.nosolousabilidad.com/articulos/descripcion_y_clasificacion.htm >. Acesso em: 16 maio 2005.

HODGE, Gail. **Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files**. The Digital Library Federation - Council on Library and Information Resources: Washington, 2000**.** Disponível em: **<** http://www.clir.org/pubs/reports/pub91/contents.html >. Acesso em: 28 fev. 2005.

HUDON, Michèle. True and tested products: thesauri on the Web. **The indexer,** Londres, v. 23, n. 3. April 2003. p. 115-119.

JOHNSON, Eric H. Distributed Thesaurus Web Services. **Cataloging & Classification Quarterly**. V. 37. n. ¾, 2004, p. 121.153. DOI:10.1300/J104v37n03_09.

KNAPP, Sara D.; COHEN, Laura B. & JUEDES, D. R. A natural language thesaurus for the humanities: the need for a database search aid. **Library Quarterly**. v. 68, n.4, 1998, p. 406-430.

LANCASTER, Frederick Wilfrid. **El control del vocabulario en la recuperación de información**. 2ª ed. Saragossa: Universitat de Valencia, 2002.

LANCASTER, Frederick Wilfrid. **Information Retrieval Systems: Characteristics, Testing and Evaluation**. 2ª Ed. New York: John Wiley & Sons: 1979.

LÓPEZ ALONSO, Miguel-Ángel & MOREIRO GONZÁLEZ, José Antonio. **Presente y futuro de los thesauros como herramienta conceptual de precisión para la recuperación de la información**. Disponível em: < http://161.116.140.71/pub/fburg/docs/lopez-moreiro.pdf > Acesso em: 20 dez. 2004.

MANDALA, Rila; TOKUNAGA, Takenobu & TANAKA, Hozumi. Query expansion using heterogeneous thesauri. **Information Processing and Management**. v. 36, 2000, p. 361-378.

MASSE, Claudine & MÉNILLET, Dominique. **Thesaurus en ligne et nouveaux usages**. LES RENCONTRES DES PROFESSIONNELS DE L'IST L'IST. Jun. 2004. Disponível em: < http://rpist.inist.fr/2004/pdf/interventions/15juinatelier6i.PDF >. Acesso em: 20 jan. 2005.

MÉNDEZ RODRIGUEZ, Eva Mª. Metadados y Tesauros: aplicación de XML/RDF a los sistemas de organización del conhecimento en Intranets. **FESABID**, 2000. Disponível em: < http://www.bib.uc3m.es/~mendez/publicaciones/fesabid00/fesabid002.pdf >. Acesso em: 31 mar. 2005.

MILES, Alistair & MATTHEWS, Brian. **Review of RDF Thesaurus Work**. 2001. Disponível em: < http://www.w3c.rl.ac.uk/SWAD/deliverables/8.2.html >. Acesso em: 14 abr. 2005.

MILES, Alistair & ROGGERS, Nikki. **Use Cases**. 2004. Disponível em: < http://www.w3.org/2001/sw/Europe/reports/thes/usecase.html >. Acesso em: 07 abr. 2005.

MILES, Alistair & ROGERS, Nikki & BECKETT, Dave. **SKOS Core 1.0 Guide**. 2004. Disponível em: < http://www.w3.org/2001/sw/Europe/reports/thes/1.0/guide/ >. Acesso em: 14 abr. 2005.

MILLER, Paul. I say what I mean, but do I mean what I say? **Ariadne**. Issue 23. Disponível em: < http://www.ariadne.ac.uk/issue23/metadata/intro.html >. Acesso em: 05 mar. 2005.

MILSTEAD, Jessica L. **Use of Thesauri in the Full-Text Environment**. 1998. Disponível em: < http://www.bayside-indexing.com/Milstead/useof.htm >. Acesso em: 20 dez. 2004.

MOREIRA, Alexandra; ALVARENGA, Lídia & OLIVEIRA, Alcione de Paiva. O nível do conhecimento e os instrumentos de representação: tesauros e ontologias. **DataGramaZero - Revista de Ciência da Informação,** v.5, n.6, dez 2004. Disponível em: < http://www.dgz.org.br/dez04/Art_01.htm >. Acesso em: 19 abr. 2005.

NATIONAL INFORMATION STANDARDS ORGANIZATION (U.S.). **Guidelines for the construction, format, and management of monolingual thesauri / developed by the National Information Standards Organization: approved August 28, 2003, by the American National Standards Institute - ANSI/NISO Z39.19 – 2003 (revision of Z39.19 – 1980)**. Bethesda (USA): NISO Press, 2003. ISBN 1-880124-04-1.

NATIONAL INFORMATION STANDARDS ORGANIZATION (U.S.). Guidelines for the construction, format and management of monolingual controlled vocabularies / **developed by the National Information Standards Organization: Ballot Period: April 11 – May 25, 2005. ANSI/NISO Z39.19-200X.** Bethesda (USA): NISO Press, 2005. ISBN: 1-880124-65-3

NAUMIS, Catalina Peña. El tesauro en el ambiente digital. **Investigación Bibliotecológica: archivonomía, bibliotecología e información,** v.15, n. 31, jul./dez. 2001.

PEÑAS, Anselmo; VERDEJO, Felisa & GONZALO, Julio. Terminology Retrieval: Towards a Synergy between Thesaurus and Free Text Searching. 8TH IBERO-AMERICAN CONFERENCE ON AI: ADVANCES IN ARTIFICIAL INTELLIGENCE. **Proceedings,** 2002, p. 684-693. Disponível em: < http://nlp.uned.es/docs/iberamia2002.pdf >. Acesso em: 03 jan. 2005.

PÉREZ AGÜERA, José Ramón. **Automatización de tesauros y su utilización en la Web Semántica**. 2004. Disponível em: < http://www.w3.org/2001/sw/Europe/events/200406-esp/trabajo-final-extratesauros/trabajo-final-extratesauros.html >. Acesso em: 20 abr. 2005.

PIERRET, Jean Dominique; DOLFI, Fabrizio; QUONIAM, Luc; BOUTIN, Eric & RICCIO, Edson Luiz. **Découverte de connaissances dans les bases de données bibliographiques: Modèles expérimentaux autour de la première hypothèse de Swanson**. 2005. Disponível em: < http://archivesic.ccsd.cnrs.fr/documents/archives0/00/00/13/83/sic_00001383_00/sic_00001383.pdf >. Acesso em: 04 abr. 2005.

PINCEMIN, Bénédicte. Thésaurus documentaires et ontologies; Divergences et ressemblances. **Communication** à LA JOURNÉE D'ÉTUDE WEB SÉMANTIQUE ORGANISÉE PAR L'UNIv. Paris V et l'ADBS le 14 octobre 2003 à Paris–La Défense. Disponível em: < http://www-lli.univ-paris13.fr/membres/biblio/1195_pincemin_ws_0410.pdf >. Acesso em: 24 fev. 2005.

RESOURCE DESCRIPTION FRAMEWORK (RDF). Disponível em: < http://www.w3.org/RDF/ >. Acesso em: 10 abr. 2005.

ROSENFELD, Louis & MORVILLE, Peter. **Information Architecture for the World Wide Web**. 2ªed. Sebastopol (EUA): O´Reilly, 2002.

SAJUS, Bertrand. **La fonction thésaurale au coeur des systèmes d´information**. 2002. Disponível em: < http://www.adbs.fr/uploads/journees/2315_fr.php >. Acesso em: 17 mar. 2005.

SANTOS, Nilton Bahlis dos. **A Ciência da Informação e o Paradigma Holográfico: A Utopia de Vannevar Bush**. Orientador: Aldo de Albuquerque Barreto, PHD. Rio de Janeiro, 2005. 185 p. Tese de Doutoramento em Ciência da Informação. Rio de Janeiro: IBICT/ECO. Orientador Aldo de Albuquerque Barreto. Disponível em: < http://biblioteca.ibict.br/phl8/anexos/niltonsantos2005.pdf >. Acesso em: 29 mar. 2005.

SILVEIRA, Maria L. & RIBEIRO NETO, Berthier. Concept-based ranking: a case study in juridical domain. **Information processing and management**, v. 40, 2004, p. 791-805.

SHIRI, Ali Asghar & REVIE, Crawford. Thesauri on the Web: current developments and trends. **Online Information Review**. v. 24, n.4, 2000, p. 273-279.

SHIRI, Ali Asghar; REVIE, Crawford & CHOWDHURY, Gobinda. Thesaurus-enhanced search interfaces. **Journal of Information Science**, v. 28, n.2, 2002, p. 111-122.

SOERGEL, Dagobert. **Functions of a thesaurus / classification / ontological knowledge base**. College of Library and Information Services. University of Maryland. Oct. 1997. Disponível em: < http://www.clis.umd.edu/faculty/soergel/soergelfctclass.pdf >. Acesso em: 13 jan. 2005.

_____. Thesauri and ontologies in digital libraries: tutorial. EUROPEAN CONFERENCE ON DIGITAL LIBRARIES (**ECDL 2002**). Rome, Italy. September 15, 2002. Disponível em: < http://www.dsoergel.com/cv/B63_rome.pdf >. Acesso em: 15 abr. 2005.

SOERGEL, Dagobert; LAUSER, Boris; LIANG, Anita; FISSENA, Frehiwot; KEIZER, Johannes & KATZ, Stephen. Reengineering Thesauri for New Applications: the AGROVOC Example. **Journal of Digital Information**, v.4, Issue 4, article n.257, mar. 2004. Disponível em: < http://jodi.ecs.soton.ac.uk/Articles/v04/i04/Soergel/ >. Acesso em: 16 mar. 2005.

ZAVITOSKI, Maria Teresa. **Exploração do uso do tesauro como instrumento de recuperação da informação**. São Paulo, 2001. Dissertação (Mestrado) – Escola de Comunicações e Artes, Universidade de São Paulo.

Bibliografia complementar

ABDALLA, Eidi Raquel Franco. **A seleção da metodologia da pesquisa por mestrandos em Biblioteconomia e Ciência da Informação**. 2003. Dissertação (Mestrado em Ciência da Informação) – Escola de Comunicações e Artes, Universidade de São Paulo, São Paulo, 2003.

DODEBEI, Vera Lucia Doyle. **Tesauro: linguagem de representação da memória documentária**. Niterói: Intexto; Rio de Janeiro: Interciência, 2002.

HAN-QING, Hou & CHUN-XIANG, Xue. **Construction of Knowledge Base for Automatic Indexing and Classification Based on *Chinese Library Classification*.** Disponível em: <

http://www.fao.org/agris/aos/ConferencesW/FifthAOS_China04/AOS_Proceedings/docs/3-3.pdf > Acesso em: 25 abr. 2005.

HUNG, Nguyen Manh. Thesaurus implementation in integrated system of information resources (ISIR). **Programming and Computer Software**, v. 30, n. 4, 2004, p. 230–240.

IBICT. **Diretrizes para o estabelecimento e desenvolvimento de Tesauros Monolíngües**. Brasília: IBICT/SENAI, 1993.

JESUS, Jerocir Botelho Marques de. Tesauro: Um instrumento de representação do conhecimento em sistemas de recuperação da informação. XII SEMINÁRIO NACIONAL DE BIBLIOTECAS UNIVERSITÁRIAS. **Anais**. Recife: UFPE, 2002. Disponível em: < http://www.ndc.uff.br/textos/jerocir_tesauros.pdf >. Acesso em: 17 dez. 2004.

JING, Yufeng & CROFT, W. Bruce. An Association Thesaurus for Information Retrieval. 4TH INTERNATIONAL CONFERENCE “RECHERCHE D'INFORMATION ASSISTEE PAR ORDINATEUR". **Proceedings** of RIAO-94. Disponível em: < http://citeseer.ist.psu.edu/jing94association.html >. Acesso em: 29 dez. 2004.

JOHNSON, Eric H. & COCHRANE, Pauline A. A Hypertextual Interface for a Searcher's Thesaurus. DIGITAL LIBRARIES '95 - THE SECOND ANNUAL CONFERENCE ON THE THEORY AND PRACTICE OF DIGITAL LIBRARIES. **Proceedings**. 1995. Disponível em: < http://www.csdl.tamu.edu/DL95/papers/johncoch/johncoch.html > . Acesso em 14 dez. 2004.

KAJI, Hiroyuki; MORIMOTO, Yasutsugu; AIZONO, Toshiko & YAMAZAKI, Noriyuki. **Corpus-dependent Association Thesauri for Information Retrieval**. 2000. Disponível em: < http://portal.acm.org/citation.cfm?id=990879 >. Acesso em: 13 jan. 2005.

LIMA, Vânia Mara Alves. **Terminologia, Comunicação e Representação Documentária**. São Paulo, 1998. Dissertação (Mestrado) – Escola de Comunicações e Artes, Universidade de São Paulo.

TRISTÃO, Ana Maria Delazari; FACHIN, Gleisy Regina Bóries & ALARCON, Orestes Estevam. Sistema de classificação facetada e tesauros: instrumentos para a organização do conhecimento. **Ci. Inf.**, Brasília, v. 33, n. 2, maio/ago. 2004, p. 161-171.

YANG, Christopher C. & LUK, Johnny. Automatic Generation of English/Chinese Thesaurus Based on a Parallel Corpus in Laws. **Journal Of The American Society For Information Science And Technology**, v.54, n.7, 2003, p.671–682.

ZAZO, Ángel F.; FIGUEROLA, Carlos G.; ALONSO, José L. Berrocal & RODRÍGUEZ, Emilio. Reformulation of queries using similarity thesauri. **Information Processing and Management** xxx (2004) xxx–xxx. DOI:10.1016/j.ipm.2004.05.006.

World Wide Web Consortium. **Architecture of the World Wide Web, Volume One**. 2004. Disponível em: < http://www.w3.org/TR/webarch/ >. Acesso em: 25 abr. 2005.

ANEXO A – Tradução das estruturas dos tesauros na ANSI/NISO Z39.19-2003

Estrutura dos tesauros

Os tópicos abaixo são uma tradução de partes da norma ANSI/NISO Z39.19-2003, feita apenas como propósito acadêmico.

Componentes do tesauro

Os componentes dos tesauros são descritos na norma ANSI/LLNISO Z39.19-2003 da seguinte maneira:

1- Descritores

Cada descritor incluído em um tesauro representa um conceito único (ou unidade de idéia). Um conceito pode ser expresso por um termo de única palavra ou por um termo com múltiplas palavras.

O escopo dos descritores é restringido para selecionados significados em um domínio do tesauro. Cada descritor pode ser formulado em tal modo que convir o escopo intencionado para cada usuário do tesauro.

Os conceitos representados por descritores podem ser agrupados em tipos gerais. Os tipos listados não são exaustivos:

a) coisas e suas partes físicas
b) materiais
c) atividades ou processos
d) eventos ou acontecimentos
e) propriedades ou condições de pessoas, coisas, materiais ou ações
f) disciplinas ou campos de assuntos
g) unidades de medida

1.1 - Qualificador Parentético *(Parenthetical Qualifiers)*

O uso de termos homográficos como descritores requer clarificação dos seus significados por meio de qualificadores (‘gloss’[29] em terminologia lingüística). O qualificador, que está cercado entre parênteses, é uma parte do descritor. O próprio qualificador pode ser um descritor, freqüentemente um termo geral que o deve ser qualificado. Ele precisa ser o mais conciso possível, idealmente consistido de uma palavra, mas não pode ser uma homográfica. Qualificadores podem ser padronizados em um dado tesauro para possível amplitude.

Um qualificador não é uma nota de escopo; contudo, um descritor qualificado pode ter uma nota de escopo apensada a ele.

Qualificadores podem também ser adicionados para termos de entrada (*entry terms*) quando seu significado é ambíguo.

1.2 - Notas de escopo *(Scope Note)*

As notas de escopo são usadas para restringir ou expandir a aplicação de um descritor, para distinguir entre descritores que têm significados sobrepostos na linguagem natural, ou para prover conselho no uso de outro termo para o indexador e o pesquisador. Uma nota de escopo deve situar o significado escolhido de um descritor; ele pode também indicar outros significados que são reconhecidos em linguagem natural, mas que foram deliberadamente excluídos do vocabulário controlado. Uma nota de escopo, ao contrário de um qualificador parentético, não é uma parte de um descritor. Embora qualificadores são geralmente adicionados somente como homógrafos, uma nota de escopo (**SN**, em inglês) pode ser fornecida para cada descritor.

1.2.1 - Notas de escopo recíprocas *(Reciprocal Scope Notes)*

Quando a referencia é feita para outros descritores em uma nota de escopo, uma nota de escopo recíproca precisa geralmente ser produzida para cada descritor mencionado.

[29] Termo em inglês. Não foi possível traduzi-lo.

Mesmo quando o escopo de somente um dos descritores requer clarificação, é conveniente anotar no registro do termo para o segundo descritor que foi citado na nota de escopo de um descritor diferente.

O **X** indica que existe uma referencia na nota de escopo de um termo para o outro.

Essa referencia recíproca ira assegurar que quando uma mudança é feita para um dos descritores, ou ele é excluído, o efeito no outro descritor será considerado.

2 - Relacionamentos

Uma das principais diferenças entre tesauros e outras listas de vocabulário é forma de apresentação do tesauro e clara diferenciação os relacionamentos semânticos básicos que liga os termos que eles contém por meio de indicadores de relacionamentos. Outros tipos de listas de termos não provêm indicadores de relacionamentos como esses, ou não provêm eles sistematicamente. De maneira ideal, um tesauro não deve incluir nenhum termo órfão, descritores que não tem relacionamentos para nenhum outro descritor.

2.1 Tipos de relacionamentos

Relacionamentos de três tipos podem ser incluídos nos tesauros:

a) Relacionamento de equivalência;
b) Relacionamento hierárquico;
c) Relacionamento associativo.

Cada relacionamento possui a propriedade de reciprocidade[30]. Os indicadores de relacionamentos são operadores binários (ou pares). Alguns indicadores são simétricos, e alguns são assimétricos:

RT é simétrico: se A RT B, então B RT A.

USE e UF são assimétricos: se A USE B, então B UF A.

Igualmente para BT e NT: se A BT B, então B NT A.

[30] Todo relacionamento indicado entre os termos A e B tem relacionamento correspondente a partir do termo B para o termo A.

Relacionamento	Indicador de relacionamento	Abreviação
Equivalência (Sinonímia)	USE USED FOR	**Nada ou U** **UF**
Hierarquia	BROADER TERM NARROWER TERM	**BT** **NT**
Associação	RELATED TERM	**RT**

Quadro 2 : Abreviações convencionais para indicadores de relacionamentos

Fonte: ANSI/NISO Z39.19-2003

A lista completa é:

Abreviações (Códigos do Tesauro) e indicadores de relacionamentos

BT = broader term

BTG = broader term (generic)

BTI = broader term (instance)

BTP = broader term (partitive)

GS = generic structure

HN = history note

NT = narrower term

NTG = narrower term (generic)

NTI = narrower term (instance)

NTP = narrower term (partitive)

RT = related term

SEE = equivalent to U (USE)

SN = scope note

TT = top term

U = use

UF = used for

UF+ = used for . . . and . . .

USE+ = use . . . and . . .

X = see from (equivalent to UF); reciprocal of *see*

2.1.1 - Relacionamentos de equivalência

Quando um mesmo conceito pode ser expresso por dois ou mais termos, um deles é selecionado como termo preferido (descritor). A relação entre termos preferidos e não preferidos é uma relação de equivalência em que cada termo é considerado como referindo para o mesmo conceito. O descritor realmente substitui para outros termos expressando conceitos equivalentes ou aproximadamente equivalentes. Uma referencia cruzada para o descritor pode ser feita a partir de qualquer sinônimo ou quase sinônimo que pode funcionar como um termo de entrada para o usuário.

Reciprocidade das relações de equivalência é expressa pelas seguintes convenções:

U or USE Qualquer condução de um termo não preferido para um descritor, e

UF or USED FOR a recíproca, que registros de termos de entrada conduzem ao descritor.

Esses indicadores de relacionamentos são os equivalentes de ***see*** e ***x*** (*see from*), respectivamente, em muitas tradicionais listas de cabeçalhos de assunto. Essa relação de equivalência cobre três tipos básicos do termo:

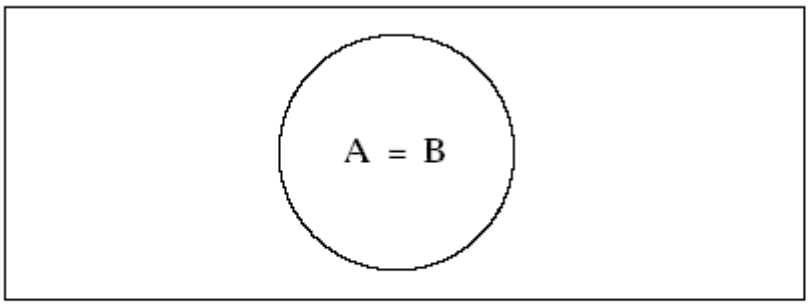
a) sinônimos;

b) variantes lexicais; e

c) quase sinônimos.

A figura abaixo representa o relacionamento de equivalência:

The circles in the diagrams below represent the scope of the terms.



The diagram implies equivalent sets. Circle A and Circle B overlap.

Figure 1a. The equivalence relationship, e.g., teenagers USE adolescents.
(A) (B)

Figura 2 : Relacionamento de equivalência
Fonte: ANSI/NISO Z39.19-2003

2.1.1.1 - Relacionamento hierárquico *(The Hierarchical Relationship)*

Esse relacionamento básico é a característica primária que distingue um tesauro sistemático de uma lista de termos desestruturada, como um glossário. Ele é baseado em graus ou níveis de superordinação ou subordinação, aonde o descritor superordinado representa uma classe ou um todo, e descritores subordinados referem se como seus membros ou partes. Reciprocidade pode ser expressa pelas seguintes indicadores de relacionamento:

BT (*Broader Term*), o nome para o descritor superordinado.

NT (*Narrower Term*), o nome para o descritor subordinado.

No formato de tesauro plano, **BT** e **NT** indicam um nível acima e um nível abaixo, respectivamente. Existem outros tipos de apresentações alfabéticas que indicam múltiplos níveis da hierarquia.
Relacionamentos hierárquicos podem também ser indicado por apresentações sistemáticas como as estruturas arbóreas ou apresentações gráficas.

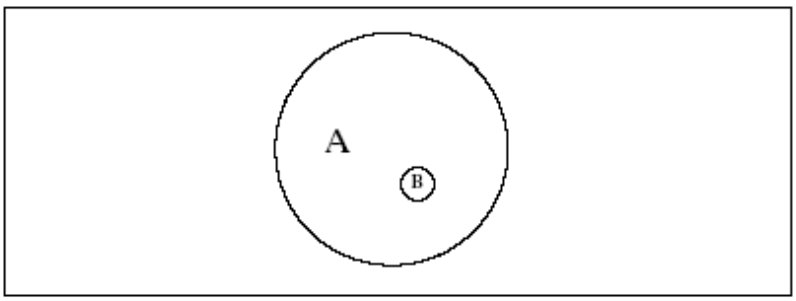
O relacionamento hierárquico cobre três situações exclusivas diferentes logicamente e reciprocamente:

a) A relação genérica (*the generic relationship*)

b) A relação todo – parte (*the whole-part relationship*);
c) a relação de exemplificação (*the instance relationship*).

Todos esse tipos de relacionamento podem dirigir-se a um único descritor, e códigos especiais podem ser usados para distinguí-los. Cada desses relacionamentos direcionam para hierarquias que são acessíveis para um teste lógico como referência para os tipos básicos de conceitos representados pelos termos. Todo descritor subordinado pode se referir para o mesmo tipo de conceito que seu descritor superordinado, que são, também o termo geral e o específico representando uma coisa, uma ação, uma propriedade, etc.

A figura abaixo representa o relacionamento hierárquico:



The diagram implies class inclusion.

Figure 1b. The hierarchical relationship, e.g., mammals NT cows.
(A) (B)

Figura 3 : Relacionamento hierárquico
Fonte: ANSI/NISO Z39.19-2003

2.1.1.1.1 - O Relacionamento genérico *(The Generic Relationship)*

Esse relacionamento identifica a ligação entre uma classe e seus membros ou espécies. Um modo simples par aplicar o teste de validação descrita acima é formular a afirmação "[*narrower term*] é um [*broader term*]".

Os códigos usados para se reconhecer esse tipo de relacionamento são:

BTG = *Broader term (generic)*

NTG = *Narrower term (generic)*

2.1.1.1.2 - O relacionamento Todo – Parte *(The Whole-Part Relationship)*

Esses relacionamentos cobrem situações em que um conceito é inerentemente incluído em outro, indiferente do contexto, de tal modo que os descritores podem ser organizados em hierarquias lógicas, com um todo tratado como um termo geral (*Broader term*). Esse relacionamento pode ser aplicado para diversos tipos de termos; os quatro tipos enumerados abaixo não têm a intenção de serem exaustivos:

a) sistemas e órgãos do corpo;
b) localizações geográficas;
c) disciplinas ou campos do conhecimento;
d) hierárquicas organizacionais, corporativas, sociais ou estruturas políticas.

Os códigos usados para se reconhecer esse tipo de relacionamento são:

BTP = *Broader term (partitive)*
NTP = *Narrower term (partitive)*

2.1.1.1.3 - O relacionamento de exemplificação *(The Instance Relationship)*

Esse relacionamento identifica uma ligação entre uma categoria geral de coisas ou eventos, expressadas por um nome comum, e um exemplo individual para essa categoria, freqüentemente um nome próprio.

Os códigos usados para se reconhecer esse tipo de relacionamento são:

BTI = *Broader term (instance)*
NTI = *Narrower term (instance)*

2.1.1.2 - O relacionamento associativo (The Associative Relationship)

Esse relacionamento cobre associações entre descritores que não são equivalentes ou hierárquicos; contudo os termos são semanticamente ou conceitualmente associados de tal modo que as ligações entre eles devem ser criadas explicitamente no tesauro, na área dele

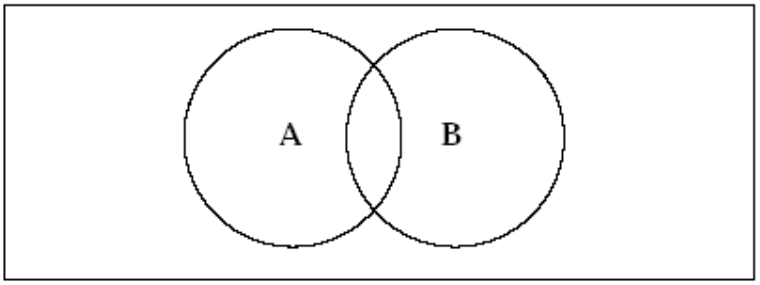
pode sugerir descritores adicionais para uso em indexação ou recuperação. O relacionamento é simétrico, e é geralmente indicado pela abreviação **RT** (*related term*).

O relacionamento associativo é um dos mais difíceis de definir, porém ele é importante para tornar explicito a natureza do relacionamento entre descritores ligados desse modo e para evitar julgamentos subjetivos o máximo possível; por outro lado, referencias **RT** podem ser estabelecidas inconsistentemente.

Como uma diretriz geral, sempre que um termo é usado, o outro deve sempre ser incluído dentro a armação comum de compartilhada relevância pelos usuários dos tesauros. Além disso, um dos termos é freqüentemente um componente necessário para qualquer explicação ou definição de outro. Qualquer um dos dois tipos de termos seguintes pode ser ligado por relação associativa:

a) aquela pertencente à mesma hierarquia;
b) aquela pertencente a diferentes hierarquias.

A figura abaixo representa o relacionamento associativo:



The diagram implies semantic overlap, i.e., that there is an element of meaning common to both terms.

Figure 1c. The associative relationship, e.g., gold RT money.
(A) (B)

Figura 4 : O relacionamento associativo
Fonte: ANSI/NISO Z39.19-2003

APÊNDICE A - Pequena comparação do esboço de revisão da norma ANSI/NISO Z39.19-200X com a norma ANSI/NISO Z39.19-2003 (vigente)

A mudança de nome de tesauro para vocabulário controlado foi feita para abranger, além de tesauros, estruturas menos complexas como listas, anel de sinônimos e taxonomia.

A diferença entre essas estruturas é a complexidade envolvida, conforme podemos observar na figura:

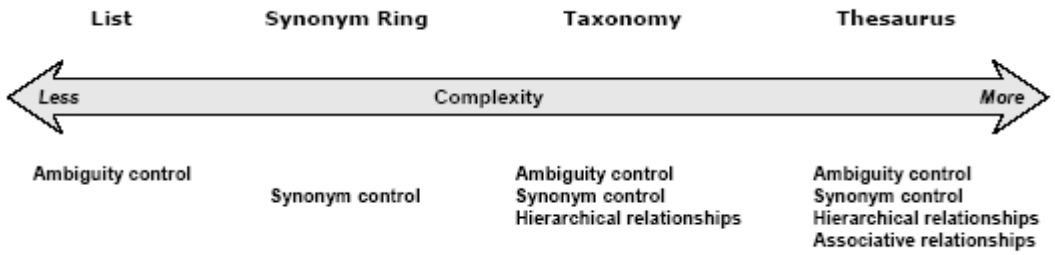


Figure 5: Increasing structural complexity among controlled vocabularies

Figura 5 : Complexidade estrutural crescente entre vocabulários controlados
Fonte: ANSI/NISO Z39.19-200X

Uma inovação da ANSI/NISO Z39.19-200X é a inclusão da Analise facetada, cuja base é o trabalho de Ranganathan.

Na composição dos termos, construíram regras mais claras para o uso de qualificadores e criaram o chamado “History Notes”[31] , que são identificadas pelo símbolo HN.

As regras para a inclusão de termos, assim como para formas gramaticais e termos compostos permitidos no tesauros se ampliaram.

[31] Notas históricas

Os relacionamentos entre os termos explicitados na norma são mais específicos em relação aos explicitados na norma ANSI/NISO Z39.19-2003:

Relationship Type	Example
Equivalency	
Synonymy	UN / United Nations
Lexical variants	pediatrics / paediatrics
Near synonymy	sea water / salt water smoothness / roughness
Hierarchy	
Generic or IsA	birds / parrots
Instance or IsA	sea / Mediterranean Sea
Whole / Part	brain / brain stem
Associative	
Cause / Effect	accident / injury
Process / Agent	velocity measurement / speedometer
Process / Counter-agent	fire / flame retardant
Action / Product	writing / publication
Action / Property	communication / communication skills
Action / Target	teaching / student
Concept or Object / Property	steel alloy / corrosion resistance
Concept or Object/ Origins	water / well
Concept or Object / Measurement Unit or Mechanism	chronometer / minute
Raw material / Product	grapes / wine
Discipline or Field / Object or Practitioner	neonatology / infant

Quadro 3 : Relacionamentos semânticos selecionados entre termos
Fonte: ANSI/NISO Z39.19-200X

Outra mudança importante é em relação às construções de apresentação de tesauros, que estão mais adaptados aos sistemas computacionais, inclusive a Web. Mas é importante notar que não são mencionados os formatos legíveis por máquinas.

Outra questão importante é a orientação sobre a interoperabilidade entre textos.