Manoel Palhares Moreira

Ambiente para geração e manutenção semi-automática de tesauros

Belo Horizonte
2005

Manoel Palhares Moreira

Ambiente para geração e manutenção semi-automática de tesauros

Tese apresentada ao Programa de Pós-Graduação em Ciência da Informação da Escola de Ciência da Informação da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do Grau de Doutor em Ciência da Informação.

Linha de Pesquisa: Tratamento da Informação

Orientadora: Profª. Dra. Maria Aparecida Moura

Belo Horizonte
Escola de Ciência da Informação da UFMG
2005

Moreira, Manoel Palhares.
M836a Ambiente para geração e manutenção semi-automática de tesauros [manuscrito] / Manoel Palhares Moreira. – 2005.
197 f. : il.

Orientadora: Maria Aparecida Moura.
Tese (doutorado) – Universidade Federal de Minas Gerais, Escola de Ciência da Informação.
Referências: f. 168-178
Anexos: f. 179-197

1. Ciência da informação - Teses 2. Tesauro – Teses 3. Indexação automática – Teses 4. Tecnologia da informação – Teses 5. Sistemas de recuperação da informação – Tecnologia – Teses I. Título II. Moura, Maria Aparecida III. Universidade Federal de Minas Gerais. Escola de Ciência da Informação.

CDU: 025.4.03:004.4

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

UFMG

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

FOLHA DE APROVAÇÃO

"AMBIENTE PARA GERAÇÃO E MANUTENÇÃO SEMIAUTOMÁTICA DE TESAUROS"

Manoel Palhares Moreira

Tese submetida à Banca Examinadora, designada pelo Colegiado do Programa de Pós-Graduação em Ciência da Informação da Universidade Federal de Minas Gerais, como parte dos requisitos à obtenção do título de **"Doutor em Ciência da Informação"**, linha de pesquisa **"Organização e Uso da Informação (OUI)"**.

Tese aprovada em: 20 de dezembro de 2005.

Por:

Profa. Dra. Maria Aparecida Moura – ECI/UFMG (Orientadora)

Profa. Dra. Beatriz Valadares Cendón - ECI/UFMG

Profa. Dra. Lígia Maria Arruda Café – UFSC

Prof. Dr. Sílvio Jamil Ferzoli Guimarães – PUC/MG

Prof. Dr. Eduardo José Wense Dias – ECI/UFMG

Profa. Dra. Gercina Ângela Borém de Oliveira Lima – ECI/UFMG

Prof. Dr. Renato Rocha Souza – ECI/UFMG

Aprovada pelo Colegiado do PPGCI

Profa. Marlene de Oliveira
Coordenadora do PPGCI/UFMG

Versão final Aprovada por

Profa. Maria Aparecida Moura
Orientadora

Meus pais se foram durante o tempo de doutorado.
Eram pessoas simples, com pouco estudo e felizes...
Ensinaram-me da vida muito mais que qualquer escola.
Por tudo, este trabalho é dedicado a eles.

Agradecimentos

Muitas coisas aprendi durante o tempo de doutorado; muito da ciência, muito mais das pessoas. Antes de terminá-lo, preciso agradecer a todos aqueles que facilitaram este percurso.

Agradecer a meus pais, que sempre estarão conosco e que na simplicidade de suas vidas souberam nos dar oportunidades que nunca tiveram. A eles devo tudo. O que me ensinaram permanecerá.

Agradecer a minha orientadora Profª. Maria Aparecida Moura, pela dedicação, tranqüilidade e firmeza na condução deste trabalho; agradecer por me ensinar da vida através da força e da energia que sabe transmitir.

Agradecer à Profª. Beatriz Valadares Céndon por ter me acolhido no programa de Ciência da Informação, pela atenção, pelo carinho e as orientações durante o percurso.

Agradecer a minha eterna professora Alcenir Soares dos Reis, pela amizade que temos, por me escutar nos momentos mais difíceis e ter sempre algo para dizer e me fazer pensar.

Agradecer aos professores Eduardo José Wense Dias, Gercina Ângela Borem Oliveira Lima e Silvio Jamil Ferzoli Guimarães, que participaram da banca de qualificação deste projeto e muito contribuíram com opiniões e sugestões para o mesmo.

Agradecer a Profª. Maria Helena de Andrade Magalhães pela atenção e gentileza na revisão minuciosa deste texto.

Agradecer aos alunos Julia Aparecida Gonçalves, Sergio Murilo Stempliuc e Carlos Henrique Palhares Moreira Junior, pelas contribuições na construção do projeto e a presteza em todo esse processo.

Agradecer aos colegas da PUCMinas - Almeida, Adriane, Ana Maria, Cristina, Goretti, Humberto, Marco Túlio, Mark, Maurício (UFMG) e Nelma - pelo apoio recebido em instantes menos fáceis que a vida trouxe junto com o doutorado e

que sem dúvida foram suavizados pela amizade que cada um deles demonstrou. Agradecer aos bibliotecários da PUCMinas - Cássio, Jane, Helenice (Lelé) e à Fátima (PUCBarreiro) - pelo profissionalismo com que conduzem nosso tempo na biblioteca e pela paciência e disponibilidade em nos ajudar sempre. Sem eles esse caminho teria sido mais difícil.

Agradecer a meus irmãos Ica, Gladys, Pedrinho e Andréa, às minhas cunhadas Lúcia Helena e Vanuza, por que nesse tempo soubemos e aprendemos mais estar juntos, e por que juntos gostamos e sabemos ser uma família. A meus sobrinhos, que trazem alegria e agitação à minha vida, e às minhas tias Coca e Dirce sempre atentas às minhas necessidades.

Agradecer a pessoas muito especiais, como a Adriana, pelo companheirismo e a cumplicidade destes anos todos de amizade; como o Edmundo e o João, meus companheiros de bons e menos bons momentos; como o Renato, pelo carinho, a atenção e a preocupação constante comigo. São todos grandes amigos que fazem sempre minha vida mais leve.

Dizer obrigado a tudo isso é muito pouco, mas é o que se sabe e o que é possível exprimir em instantes assim.

Há que se escrever a vida em flauta e
vôo como cantam os pássaros.
Buscar na memória a lembrança e a
direção. Ocultar os rastros
percorridos para perder-se no
encontro e ninho. Decifrar o alfabeto
rabiscado nas linhas do vento,
gravado no fruto maduro,
embaraçado na pena trocada. Como
os pássaros, há de se escrever
enquanto é dia e para todos.

Bartolomeu Campos Queirós

Resumo

Entre as diversas formas de representação da informação utilizadas por Sistemas de Recuperação de Informação encontram-se os tesauros, que se constituem em uma linguagem de indexação consolidada e empregada por profissionais que exercem atividades de organização da informação. A flexibilidade para o estabelecimento de novas relações entre termos, as hierarquias e as referências cruzadas conferem ao instrumento uma multiplicidade de usos, abrangendo processos que vão desde a indexação até a efetiva recuperação dos documentos. A elaboração e manutenção de tesauros são atividades intelectuais com procedimentos específicos, entre eles o conhecimento de documentos produzidos na área, o entendimento dos termos empregados e a construção de conceitos para explicação desses termos. Do profissional envolvido espera-se uma atitude flexível para incorporar as mudanças e inovações que surgem na área, na própria linguagem e no emprego de termos. Este trabalho objetivou a construção de uma metodologia e um ambiente para a geração e manutenção de tesauros de forma semi-automatizada, através da utilização da linguagem natural e com base em tecnologias da Ciência da Computação e nos fundamentos teóricos da Ciência da Informação; mais especificamente, através dos conceitos ordenadores da garantia literária, da garantia de uso e da garantia estrutural, incorporando-se a essas a proposta da garantia advinda da própria estrutura do texto. O ambiente possibilitou a verificação da atualidade e do potencial representativo do tesauro. Partiu-se da hipótese de que palavras-chaves recolhidas de artigos científicos poderiam ser aplicadas neste processo já que elas representam duplamente a garantia literária e a de uso por se tratar de um instrumento privilegiado de disseminação do conhecimento científico. Foram feitos cálculos estatísticos envolvendo freqüência e escore padronizado nas observações de freqüência de palavras-chave no título, no resumo, no texto e na bibliografia dos artigos. Para testes, foram utilizados textos científicos dos periódicos eletrônicos Datagrama Zero e Ciência da Informação, já consolidados na área. A inexistência de um tesauro atualizado na área levou a construção de um Tesauro em Ciência da Informação (TCI), a partir de tesauros existentes em português (IBICT), em inglês (ASIS) e em espanhol (CINDOC e DOCUTES). O ambiente apontou a necessidade de atualização de termos, classificados por grau de relevância, levando em conta a evolução da área.

Palavras-chave: tesauro; linguagem de indexação; organização da informação; automação de tesauros; mineração de palavras

Abstract

Thesauri are among the diverse means of representing information as used by Information Retrieval Systems, which is considered to be a consolidated indexing language employed by professionals that carry out activities of organizing information. The flexibility for establishing new relations between terms, the hierarchies and the crossed references give that instrument a diversity of usage, reaching processes that range from indexing to effective recovery of documents. The production and maintenance of a thesaurus are intellectual activities with specific procedures to be followed. Among them are knowledge about documents produced in the subject area, comprehension of the used terms and the construction of concepts to explain those terms. It is expected from the professionals in this field a flexible attitude to assimilate the changes and innovations which may be found in the indexing area, in the language itself, and in the usage of terms. This study aimed at the construction of a methodology and an environment to the generation and maintenance of a thesaurus within a semi-automatic way, through the use of natural language based on the technology of Computer Science and the theoretical scope of Information Science. Moreover, through the ordering concepts of literary, usage and structural guarantee, they incorporate the proposal of the guarantee that comes from the structure of the text itself. The environment made it possible to verify the present thesaurus as well as its representative potential. The hypothesis was that key words from scientific articles could be applied in this process, since they represent both the literary and usage guaranties, for they are a privileged instrument in disseminating the scientific knowledge. Statistic calculi were made involving frequency and score standardised in the observation of the frequency of key words in titles, summaries, texts and articles of the reference list. Scientific texts of the electronic periodicals *Datagrama Zero* and *Ciência da Informação* were used. The absence of an updated thesaurus in the area led to the elaboration of a Thesaurus in Information Science (TIS), from the existing thesaurus in Portuguese (IBICT), in the English language (ASIS) and in Spanish (CINDOC and DOCUTES). The environment pointed out the need of updating terms classified by degrees of relevance, considering the progress of the area.

Key-words: thesaurus; indexation language; information organization; automatic thesaurus; words mining.

résumé

Parmi les multiples formes de représentation de l'information utilisées par des Systèmes de Récupération d'Information se trouvent les thésaurus qui sont constitués d´un langage d'indexation consolidé et employé par des professionnels qui s´occupent de l´organisation de l'information. La flexibilité à établir de noveaux rapports entre les termes, les hiérarchies et les références croisées confèrent à cet outil une multiplicité d'usages comprenant des procédés dès l'indexation jusqu'à l'effective récupération des documents.

L'élaboration et le maintien de thésaurus sont des activités intellectuelles qui exigent des procédés spécifiques comme la connaissance des documents produits dans le domaine, la compréhension des termes employés et la construction de concepts pour mieux expliquer ces termes. De cette façon, on attend une attitude flexible du professionnel engagé à fin d´incorporer aux thésaurus les changements et les innovations du domaine, du langage même et de l'emploi de termes. Ce travail de recherche a eu pour but la construction d'une méthodologie et d'un environment appropriée à la génération et le maintien de thésaurus de façon semi-automatisée à travers l'utilisation du langage naturel et basé sur des technologies de l'Informatique et sur les fondements théoriques de la Science de l'Information à savoir les concepts ordonnateurs de la garantie littéraire, d´usage et de la garantie structurale, en incorporant la garantie issue de la structure même du texte. L´environmente a même permis de vérifier l'actualité et la portée représentative du thésaurus. On est parti de l'hypothèse selon laquelle les mots clés extraits d'articles scientifiques, à la fois garanties littéraire et de l'usage, pourraient être appliqués à ce processus parce qu'il s'agit d'un outil privilégié de dissémination de la connaissance scientifique. Des calculs statistiques comprenant la fréquence et le *score* standardisé dans les observations de mots clés dans le titre, dans le résumé, dans le texte et dans la bibliographie des articles. Pour les tests, ont été utilisés des textes scientifiques des publications électroniques *Datagrama Zero* et *Ciência da Informação*, déjà consolidées dans le domaine. L'inexistance d'un thésaurus mis à jour dans le domaine a mené à la construction d'un Thésaurus en Sciences de L'information (TCI), à partir de thésaurus en portugais (IBICT), en anglais (ASIS) et en espagnol (CINDOC et DOCUTES). L´environment a montré le besoin d'actualisation de termes classés par niveau d'importance, en considérant l'évolution du domaine.

Les mot-clefs: le thesaurus; langue de l'indexation; l'organisation de l'information; le thesaurus automatique; l'exploitation minière des mots.

Resumén

Entre las diferentes formas de representación de la información practicadas por los Sistemas de Recuperación de Información se encuentran los tesauros, que forman un lenguaje de indexación consolidada y empleada por profesionales que ejercen actividades de organización de la información. La flexibilidad para el establecimiento de nuevas relaciones entre términos, las jerarquías y las referencias cruzadas confieren al instrumento una multiplicidad de usos, abarcando procesos que van desde la indexación hasta la efectiva recuperación de los documentos. La elaboración y mantenimiento de tesauros son actividades intelectuales con procedimientos específicos, entre ellos el conocimientos de documentos producidos en la área, el entendimiento de los términos empleados y la construcción de conceptos para explicación de estos términos. Acerca del profesional envolvido se espera una actitud flexible para incorporar los cambios e innovaciones que emergen del área, del lenguaje y del uso de los términos. Este trabajo planteó la construcción de una metodología y un ambiente para la gestación y manutención de tesauros de forma semiautomática, a través de la práctica del lenguaje natural y con base en tecnologías de la Ciencia de Ordenadores y en los fundamentos teóricos de la Ciencia de Información; específicamente, mediante los conceptos ordenados de la garantía literaria, de la garantía del uso y de la garantía estructural, incorporándose a ellas la propuesta de la garantía sacada de la misma estructura del texto. El ambiente posibilitó la verificación de la actualidad y del potencial representativo del tesauro. Se partió de la hipótesis de que palabras claves recogidas de artículos científicos podrían ser aplicadas en este proceso ya que ellas representan duplemente la garantía literaria y la del uso porque se trata de un instrumento privilegiado de difusión del conocimiento científico. Se hicieron cálculos estadísticos envolviendo frecuencia y score padrón en las observaciones de frecuencia de palabras claves en el título, en el resumen, en el texto y en la bibliografía de los artículos. Como prueba de la hipótesis fueron usados textos científicos de los periódicos electrónicos *Datagrama Zero* y *Ciencia da Informação*, ya consolidados en la área. La inexistencia de un tesauro actualizado en la área trajo la construcción de un Tesauro en Ciencia de Información (TCI), a partir de tesauros existentes en portugués (IBICT), en lengua inglesa (ASIS) y en español (CINDOC y DOCUTES). El ambiente señaló la necesidad de actualizar términos clasificados por grado de significación, teniendo en cuenta la evolución del área.

Palabras-clave: tesauro; lenguaje de indización; información; minería de texto; tesauros automatizados

Lista de Figuras

Lista de Tabelas

Lista de abreviaturas e siglas

ABC	Academia Brasileira de Ciências
ASIS	American Society for Information Science
ABNT	Associação Brasileira de Normas técnicas
CBPF	Centro Brasileiro de Pesquisas Físicas
CDD	Classificação Decimal de Dewey
CINDOC	Centro de Información y Documentación Científica
CNAM	Conservatoire National d'es Arts et Métiers
CNBB	Conferência Nacional dos Bispos do Brasil
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
CSIC	Consejo Superior de Investigaciones Científicas
Fapesp	Fundação de Amparo à Pesquisa do Estado de São Paulo
FID	Federação Internacional de Documentação
IASI	Instituto de Adaptação e Inserção na Sociedade da Informação
IBICT	Instituto Brasileiro de Informação em Ciência e Tecnologia
INTD	Instituit National d'es Techniques de la Documentation
ISO	International Organization for Standardization
KWIC	keyword-in-context
KWOC	keyword-out of-context
MAM	Museu de Arte Moderna
INTD	Institut National des Techniques de la Documentation
ISOC-DC	Base de Datos de Biblioteconomía, Documentación y Política Científica
PCB	Partido Comunista Brasileiro
SI	International System of Units
SQL	Structured Query Language
TCI	Tesauro em ciência da Informação
WEB	World Wide Web

Sumário

1 Introdução

O período entre o final da década de 40 e o início dos anos 50 foi marcado pelo aumento das formas de produção e difusão de informações e pelo desejo de sua obtenção. O mundo vivia sob a ameaça dos horrores da guerra ainda recente, mas eram marcantes o otimismo e as esperanças de construção de um mundo de paz e bem-estar advindos da vitória da democracia sobre os regimes nazifascistas (RODRIGUES, 1996). Os anos de guerra impulsionaram o trabalho e o desenvolvimento em diversas áreas e fizeram com que as pessoas estivessem constantemente preocupadas com a defesa de seu país. O desenvolvimento tecnológico era uma questão de garantia de vida da população e do próprio país; soava como um alerta mundial para a importância da pesquisa científica e a necessidade de incentivo a essa atividade. Houve avanços da tecnologia bélica, das industrias farmacêutica e aérea. O conhecimento simbolizava poder e sobrevivência e a bomba atômica foi uma prova real disto. Havia consenso de que a ciência deveria ser empregada como fator de desenvolvimento econômico e social.

O desenvolvimento tecnológico teve importância decisiva na resolução da segunda guerra mundial e afetou substancialmente o direcionamento dado à indústria, principalmente a norte-americana, levando o tema do desenvolvimento na ciência e na tecnologia a ser almejado pelas diretrizes políticas de diversos países. Nos Estados Unidos, a indústria, o governo, a comunidade científica e a população em geral concordavam quanto a essa necessidade e acreditavam que a pesquisa seria a base para a mudança e o

direcionamento necessários. Ao Estado caberia o apoio nos investimentos e à indústria sua parcela de contribuição. A pesquisa aplicada seria orientada para objetivos específicos e realizada em institutos nacionais enquanto a pesquisa tecnológica ficaria sob o patrocínio da indústria e sem interferências governamentais (GUIMARÃES, 2002).

Em decorrência, cresceu a produção científica e tornou-se claro que era preciso acelerar o processo de disseminação do conhecimento produzido. Em 1945, Vannevar Bush publicou um artigo apontando a necessidade de se tornar acessível o conhecimento produzido em ciência e tecnologia e sugerindo a aplicação de tecnologia da informação como uma solução possível para a recuperação da informação frente ao acelerado crescimento da produção científica (SARACEVIC, 1995; BUSH, 1945). Muitos outros cientistas compartilharam dessa preocupação e entusiasmaram-se com as possibilidades de soluções tecnológicas para o problema.

Naquela época, a própria conjuntura e a necessidade do desenvolvimento haviam criado uma relação estreita entre o computador e as necessidades militares, principalmente nos Estados Unidos. O computador era utilizado principalmente em cálculos balísticos e os centros de pesquisa civil recebiam financiamentos do governo americano no período de guerra e posterior a ela. Estes financiamentos eram supervisionados pelo Departamento de Pesquisa e Desenvolvimento Científico, na época dirigido por Vannevar Bush (KUMAR, 1997).

Nos anos 50, a recuperação de informação ganha força entre a comunidade científica. Nos Estados Unidos, a solução para esse problema passa a ter

apoio do governo, em um primeiro instante endereçada principalmente à explosão de informação em ciência e tecnologia e mais tarde em outras áreas do conhecimento humano.

No Brasil, os anos 50 representaram um esforço centralizado no desenvolvimento do país, para superação de problemas sociais, do atraso econômico e cultural. Esse esforço refletiu-se em mudanças de direcionamento em vários segmentos da sociedade: na Igreja, através da criação de movimentos de juventude e da criação da Conferência Nacional dos Bispos do Brasil (CNBB); na política, pela da atuação de partidos de esquerda, de correntes socialistas e, principalmente, do Partido Comunista Brasileiro (PCB); nos movimentos nacionalistas, apoiados na possibilidade de desenvolvimento independente do Brasil através do capital nacional; em diversos ramos do conhecimento, especialmente nas ciências sociais em obras como as de Gilberto Freire (*Sobrados e mocambos* – 1957), Sérgio Buarque de Holanda (*Visão do paraíso* – 1959); nas Letras através das obras de Tristão de Athayde e Antônio Candido de Melo e Souza, dedicadas à compreensão e ao estudo da literatura brasileira. E não poderia ser esquecido o romance *Grande Sertão Veredas*, de Guimarães Rosa, editado em 1956, revelando "*o sentido universal contido nas temáticas regionais brasileiras*". No campo das artes, o Brasil destaca-se pela realização da Bienal Internacional de Artes Plásticas de São Paulo (1951) e da Exposição Nacional de Arte Concretista, realizada no Museu de Arte Moderna (MAM), que, como o Museu de Arte Moderna de São Paulo (MASP) havia sido criado na segunda metade dos anos 40; a criação do Arena, responsável pela divulgação do teatro como arma política; o rádio, que até então era o meio de comunicação mais privilegiado pelo público acaba

cedendo espaço à televisão que chega ao Brasil no início de 1950 (RODRIGUES, 1996).

A indústria também foi privilegiada. Após a depressão de 1929 e até 1945 o país passou por um grande desenvolvimento industrial. No início dos anos 50, a indústria brasileira apresentava empreendimentos centrados na produção de bens perecíveis e semiduráveis, particularmente os produtos de indústrias têxteis, alimentar, de vestuário, de fumo e couro. Eram, em sua maioria, indústrias de capital exclusivamente nacional, com a gerência considerada como valor na tradição das famílias. Data desta época a entrada do capital estrangeiro e já era visível a presença norte-americana na economia do país.

O presidente Getúlio Vargas apresentava comportamentos flexíveis ao capital estrangeiro e seu segundo mandato foi marcado pelo alinhamento político do Brasil aos Estados Unidos. Como sinal desse alinhamento, em 1951, a Comissão Mista Brasil-Estados Unidos elaborou projetos para o setor energético e viário. Em oposição à abertura do capital estrangeiro nasceu um movimento nacionalista de defesa do petróleo brasileiro que pressionou as instâncias governamentais fazendo com que, em 1953, o Congresso aprovasse uma lei instituindo o monopólio estatal de exploração e de refinamento do mesmo. Juscelino Kubitschek assume a presidência em 1956 e seu governo delineou uma nova e definitiva configuração para o desenvolvimento industrial do país, marcada pela presença do capital estrangeiro, principalmente nas indústrias pesada, automobilística, de materiais elétricos e eletrônicos, de eletrodomésticos, de produtos químicos e farmacêuticos, e de matéria plástica. Surgiram as multinacionais com domínios nesses segmentos, exercendo

influências na vida econômica, social e política do Brasil (KOSHIBA e PEREIRA, 1987).

Houve também um salto no desenvolvimento institucional da ciência no Brasil. A exemplo do modelo e do consenso norte-americano surgiram na década de 50 duas agências de fomento à pesquisa no país: o Conselho Nacional de Pesquisa (CNPq) e a Fundação de Amparo à Pesquisa do Estado de São Paulo (Fapesp), e também um laboratório de pesquisa nacional, o Centro Brasileiro de Pesquisas Físicas (CBPF). O CNPq foi criado em 1951, mas sua idéia vem desde os anos 20 através dos integrantes da Academia Brasileira de Ciências (ABC). A meta inicial e principal era a formação de recursos humanos para a pesquisa. A Fapesp, embora instituída em 1962, começou a ser pensada em 1942, quando foram montados os Fundos Universitários de Pesquisa para a Defesa Nacional, logo a após a entrada do Brasil na Segunda Guerra Mundial (GUIMARÃES, 2002; CNPq, 2005; FAPESP, 2005).

Assim, a Segunda Guerra Mundial marcou a humanidade, influenciou as escolhas e diretrizes dos países e trouxe conseqüências à vida de todos. Esta herança do período da guerra fez crescer a produção científica e representou para as bibliotecas um desafio quanto aos processos de indexação e recuperação de informação. Até então era possível organizar os documentos do acervo - em maioria quase absoluta constituídos por livros - através de instrumentos previamente construídos. O bibliotecário era o responsável pela guarda destes bens documentais, embora já almejasse ser um intermediário entre a necessidade de informação do usuário e o acervo disponível. Surgiram outros gêneros e suportes documentais, principalmente artigos de periódicos

científicos e relatórios de pesquisas, representando o início da era da especialização do conhecimento (FOSKETT, 1973; DODEBEI, 2002).

Com o intuito de proporcionar uma maior comunicação entre os cientistas, os periódicos científicos surgiram no século XVII, na Inglaterra, logo após a restauração da monarquia em 1660, como conseqüência da reunião de grupos que, durante os anos de guerra civil, reuniam-se em locais e cidades diferentes para debater questões filosóficas. Acalmados os ânimos, Londres foi escolhida como local oficial para essas reuniões que acabaram por levar à formação da Royal Society em 1662. Desde sua fundação, essa instituição preocupou-se com a questão da comunicação científica, influenciada pelas idéias de Bacon sobre a possibilidade de uma instituição científica. Eram prioridade a coleta e a análise das informações recebidas e existem registros de que seus membros percorriam países estrangeiros buscando dados sobre trabalhos neles desenvolvidos. Como a tarefa era dispendiosa e levava tempo, elegeram então como membros da Royal Society pessoas do estrangeiro, que cumpriam a tarefa de comunicar à entidade, através de cartas, os trabalhos desenvolvidos em seus países (MEADOWS, 1999). Só que a solução foi temporal e o volume de cartas era tão grande que foi preciso encontrar uma outra forma de divulgação dos trabalhos.

Em 1664, Denis de Sallo, um parisiense envolvido nessa forma de coleta e disseminação de informação, começou a articular um periódico destinado à publicação do que acontecia na Europa "*na república das letras*" (MEADOWS, 1999). Em 5 de janeiro de 1665 foi publicado o primeiro número do *Le Journal des Sçavans*, e com ele nascia o periodismo científico. Naquela época, conforme Lemos (1968) e Meadows (1999), o periódico tinha entre seus

compromissos: apresentar um catálogo dos principais livros ainda não publicados na Europa, com informações sobre seu conteúdo e sua utilidade; incluir apontamentos necrológios de celebridades da época, com bibliografia de suas obras; a divulgação de experimentos em física, química e anatomia para a explicação de fenômenos naturais, assim como a descrição de invenções de máquinas úteis ou curiosas; a divulgação de decisões dos tribunais civis e eclesiásticos e censuras de universidades, e levar aos leitores informações diversas para alento à curiosidade humana.

Em março de 1665 a Royal Society publicou sua revista *Philosophical Transactions* que, assim como o *Journal des Sçavans*, possuía cobertura ampla, embora tenham os dois tomado caminhos mais específicos com o passar do tempo. A partir daí, o crescimento da literatura científica foi exponencial, e a motivação encontra-se centrada na necessidade de disseminação e comunicação eficiente para a comunidade científica. Além de auxiliar na integração e na cooperação dos pesquisadores, contribui também para a legitimação, o reconhecimento do trabalho e a aceitação do pesquisador na própria comunidade (OLIVEIRA e NORONHA, 2005).

Desta forma, desde sua origem, o periódico científico desempenha importantes funções na área da ciência, categorizadas por Valério (1994) como registro, disseminação e instituição social. Como registro, é um meio formal de controle da qualidade da própria revista, além de ser uma fonte do saber científico e do conhecimento público. Como agente de disseminação da informação, fornece informações de interesse à comunidade científica e fomenta discussão sobre os pontos que nele são veiculados. Como instituição social atribui prestígio e reconhecimento aos autores, às instituições, aos editores e aos avaliadores.

Mas, se de um lado estão os produtores dos documentos desejosos de que seu conteúdo seja conhecido e disseminado, do outro lado estão os usuários, com toda diversidade de necessidades. As unidades de informação então se situam entre eles, como veículo apropriado para a disseminação de informações. E para que este encontro ocorra com eficácia, processos de indexação e recuperação de informação foram construídos ao longo dos anos, com esforços no intuito de aperfeiçoar o desempenho dos sistemas de recuperação de informação. Em linhas gerais, sistemas de recuperação de informação constituem-se do esforço humano e dos procedimentos implantados que visam facilitar a localização de informações disponíveis para os usuários, a partir das requisições feitas por estes (ARAÚJO, 1994).

Os procedimentos dizem respeito às atividades de representação, armazenamento, organização e acesso aos documentos (SALTON e MCGILL, 1983). A representação é feita através de processos de indexação, em atividade intelectual realizada por profissionais especializados em documentação. Segundo Naves (2000), esses profissionais são orientados por critérios que podem estar agrupados entre objetivos e subjetivos. Os critérios objetivos dizem respeito às informações bibliográficas, que através de regras estabelecidas são retiradas dos próprios documentos. Os subjetivos são critérios intelectuais e abrangem o tratamento temático do conteúdo do documento e a análise de assuntos.

O armazenamento envolve os processos de gerenciamento dos documentos, independente da mídia em que se encontram. A organização, assim como a representação, objetiva facilitar o usuário no acesso à informação desejada.

Todo desenvolvimento tecnológico dos últimos anos tem possibilitado maior agilidade no acesso e disseminação da informação nos diversos níveis da sociedade. O desenvolvimento das telecomunicações e da computação, e depois a confluência destes, tornaram possível o que chamamos de Tecnologia da Informação (OLIVEIRA e NORONHA, 2005) que no contexto de documentos e produções científicas inaugura, com o formato digital, novo suporte para registro de informações.

Também como fruto dessa tecnologia, nos anos 90 surgiu a *World Wide Web* (*Web*) com a vocação no próprio nome de tornar-se a grande teia mundial de informações. Não raro, a Web é confundida com a Internet, que nasceu em 1969 nos Estados Unidos. Originalmente, a Internet era apenas uma rede com a tarefa de interligar laboratórios de pesquisa. Chamava-se então ARPAnet (ARPA: Advanced Research Projects Agency). Pertencia ao Departamento de Defesa norte-americano e tinha um cunho estratégico demarcado: seria uma rede com disponibilidade contínua e deveria garantir a sobrevivência das redes de comunicação e a segurança das informações em tempos de Guerra Fria. A *Web* avançou nos propósitos da Internet e trouxe consigo o conceito de hipertexto objetivando uma interface mais amigável (SOUZA e ALVARENGA, 2004).

A possibilidade do suporte digital, aliada à tecnologia da Web, permitiu que novas formas de disseminação e recuperação de informações fossem introduzidas através da disponibilidade dos documentos em rede e do desenvolvimento de sistemas de recuperação de informação disponíveis nesse ambiente, viabilizando a chamada “sociedade em rede” (CASTELLS, 1999),

onde a informação é produzida e armazenada em lugar distinto do usuário que a recupera.

A desmaterialização dos objetos informacionais abriu novas perspectivas à representação e à recuperação dos documentos. Essas tecnologias alteraram também a forma de comunicação entre os homens e trouxeram mudanças na comunicação científica; contudo, sem perder o caminho já percorrido, devem ao mesmo tempo replicar e simular as formas já empreendidas na recuperação da informação.

Se hoje nos é dado o acesso facilitado ao conhecimento, honra seja feita à linguagem, pois é através dela que, principalmente, nos comunicamos com o grupo no qual estamos inseridos, com o qual trabalhamos ou nos dedicamos ao estudo. "*É a língua que nos constrói e é por meio dela que construímos o mundo e nossas relações com ele*" (BRUSCHINI et al., 1998). A linguagem é a expressão de nossa língua natural ou nativa, como comumente falamos. Esta linguagem torna abrangente nossa comunicação, tira-nos de nossos próprios limites.

A linguagem também é o meio pelo qual o homem cria e amplia sua consciência, pois é através dela que dar-se-ão atos como simbolizar, conceitualizar e classificar (MURRIEL, 1998). É nela que nos apoiamos para fixar nossas culturas regionais, nos identificamos como um povo de determinado lugar e em determinado tempo. Este é o poder da língua manifestado e modificado no cotidiano, em nosso modo de viver o *cultus* e vigorar nossa própria cultura.

Na Web, os sistemas de recuperação de informação, mais precisamente, as máquinas de busca utilizam-se de linguagem natural para os processos de recuperação de informações. Para Lopes (2002), a linguagem natural pode ser conceituada como *"sinônimo de discurso comum"*, ou seja, é a linguagem que uma comunidade utiliza em seu dia a dia, habitualmente na fala e na escrita. A linguagem natural é a linguagem utilizada nos textos científicos. Consideram-se linguagens de indexação os instrumentos de representação de informação para a indexação, o armazenamento e a recuperação de documentos. As mais conhecidas são os tesauros e os sistemas de classificação bibliográfica.

Os tesauros constituem uma linguagem de documentação com a característica específica de possuir relações entre os termos que o compõem. O termo linguagem de documentação compreende, genericamente, os sistemas de classificação bibliográfica, as listas de cabeçalho de assunto e os tesauros, os quais surgiram estimulados pela necessidade de manipulação de grande quantidade de documentos de conteúdos especializados.

O estudo dos tesauros envolve diversos campos do conhecimento, pois o tema é, em si, multidisciplinar. Diz respeito à Ciência da Informação, por sua origem e utilização imediata nos processos de indexação e recuperação de informação; também está vinculado às contribuições da Terminologia, da Lingüística, da Filosofia, da Lógica, da tradução, da análise sistêmica, da normalização, dos sistemas de classificação, entre tantos outros. Todos eles, cada qual em seu nicho particular, apresentam pontos de contribuição. Como área de estudo da Ciência da Informação, multidisciplinar por natureza e interdisciplinar na vocação, o tesauro herda dessa ciência essas propriedades.

Requer também o conhecimento do tema de tratamento da informação em sua forma abrangente, o que corresponde dizer que compreender as ferramentas anteriores, que facilitavam os serviços de indexação, colabora para o entendimento de como os tesauros passaram a representar um ganho nesta atividade. De forma geral, os tesauros são específicos, construídos para determinada área do conhecimento, que tendem a aprofundar, e raramente abrangem muitas disciplinas. Três pontos fundamentais - a garantia literária, a garantia de uso e a garantia estrutural - constituem o referencial para a construção de tesauros.

O presente trabalho trata da construção de um ambiente especializado na geração e manutenção semi-automática de tesauros envolvendo as áreas já mencionadas e conceitos. Baseia-se nas garantias literária, de uso e estrutural, a elas adicionando a garantia que vem da estrutura do documento. Encontra-se organizado da seguinte forma: a seguir apresenta-se o problema a ser investigado e os objetivos desta pesquisa; no segundo item faz-se uma revisão de literatura; logo após apresenta-se o instrumento desenvolvido para o experimento e a análise dos dados obtidos. E em forma de conclusão, apresentam-se a leitura deste caminho e as expectativas de sua continuidade.

1.1 O problema

Pesquisas em recuperação de informação foram amplamente aplicadas durante anos. As idéias e experiências dos anos 50 e 60 transformaram-se em bancos de dados próprios para recuperação de informação e em serviços e sistemas especializados no assunto. A chegada da tecnologia de redes nos

anos seguintes alterou a trajetória mas reforçou estas idéias. Houve desenvolvimento no setor da indústria de tecnologia da informação e muitos são os profissionais de informação envolvidos no assunto. A recuperação de informação passou a ser utilizada na Web e é o ponto central em bibliotecas digitais (SARACEVIC, 1995).

Mas permanecem alguns problemas:

- a escolha da linguagem a ser utilizada na indexação ainda é impasse e ponto de estudo: se a linguagem natural favorece os processos de indexação automática, pode aumentar a revocação e diminuir a precisão da recuperação de documentos. Por outro lado, a utilização de uma linguagem de indexação aumenta a precisão, mas pode distanciar-se da linguagem compreendida e empregada pelos usuários;

- a informação tem, muitas vezes, período de vida breve e precisa ser acessada rapidamente antes que perca seu valor para o usuário. A rapidez com que as informações sofrem obsolescência é desafio para os profissionais, frente ao grande número de documentos a serem indexados;

- não existem padrões para o desenvolvimento de sistemas de informação e muitas vezes o usuário perde grande tempo para compreender o funcionamento das rotinas que compõem o sistema. Outro problema é a forma como estes sistemas interagem com o usuário apesar dos avanços nas pesquisas sobre a interação homem-computador;

- a velocidade da produção é muitas vezes maior que a capacidade de indexação dos profissionais de informação, tornando o processo moroso

frente à necessidade de se cumprir o papel de comunicação entre autores e leitores;

- percebe-se que os sistemas automatizados poderiam permitir o acompanhamento da atualização dos conceitos empregados em linguagens de indexação, permitindo um histórico do emprego desses conceitos e termos; mas, não apresentam essa facilidade, muitas vezes pelo distanciamento do profissional que constrói estes sistemas e os profissionais de informação.

O quadro reflete carências em muitos pontos:

- os profissionais de Ciência da Informação nem sempre possuem instrumentos que traduzam suas necessidades de recuperação de informação, de acompanhamento da utilização destes sistemas pelos usuários e da evolução dos conceitos e do emprego de termos em áreas específicas do conhecimento;

- muitas vezes, os profissionais da informação não estão habilitados à utilização de recursos computacionais que podem acelerar o processo, o que os leva a um acúmulo de funções e a desvios no próprio ato de exprimir o que desejam;

- as ferramentas disponíveis não levam em conta alguns princípios básicos do processo de indexação e seus resultados não representam uma resposta às necessidades dos sistemas de recuperação.

Muito esforço tem sido empregado pela área de Ciência da Computação para produção de índices e listas que auxiliem o processo de recuperação da

informação, principalmente na Web. Em geral, a indexação de documentos na Web ocorre através da linguagem natural utilizada na produção do próprio documento sem que qualquer tratamento seja dado ao processo. A maioria dos trabalhos nesta área aprofunda-se em questões relativas ao desempenho da recuperação e ao poder de revocação com índices de precisão. Mas, estes trabalhos não são interdisciplinares, não contam com a participação de profissionais da informação que conheçam as necessidades dos usuários e nem mesmo de profissionais da própria área abrangida pelo trabalho.

Os princípios norteadores de tesauros – a garantia literária, a garantia do usuário e a garantia estrutural – também são por vezes omitidos. Se a primeira pode ser justificada nestas pesquisas pelo fato dos descritores terem sido extraídos de textos da própria área, na maioria das vezes o usuário é relegado a segundo plano, desconsiderando a forma como ele interage com os assuntos dos quais busca a informação. Também os processos que agregam facilidades para que os usuários entendam a estrutura de indexação não estão explicitados. Permanece a necessidade, para os profissionais de informação, de construir um ferramental de indexação que considere estes princípios e que utilize a tecnologia da computação para colaborar em todo o processo.

É necessário que sejam desenvolvidas ferramentas para auxílio à indexação que levem em conta a facilidade de indexação da linguagem natural, sem contudo esquecer os benefícios do tratamento da informação e das linguagens de documentação, assegurando que esse processo produza resultados que correspondam à realidade da área.

Se a geração dessas ferramentas constitui um problema a ser resolvido, manter o seu produto atualizado também é outro problema. A produção de documentos avança pelas diversas áreas da ciência e novidades na pesquisa são sempre incorporadas. Faz-se necessário que uma ferramenta para auxílio à indexação de documentos possua a capacidade de refletir e acompanhar toda esta evolução.

Os tesauros constituem uma ferramenta de indexação já consolidada nas atividades de organização da informação empregada por muitos que exercem essas atividades. A flexibilidade de estabelecimento de novas relações entre os termos de um tesauro, o estabelecimento de hierarquias e referências cruzadas conferem ao instrumento uma multiplicidade de usos, abrangendo os processos desde a indexação até o suporte para a efetiva recuperação dos documentos.

Já existem ferramentas que automatizam algumas atividades do processo de indexação e até algumas mais específicas que colaboram com a construção de tesauros. Como exemplo, as máquinas de busca na Web, as pesquisas na computação para acelerar o processo de busca em textos, de criação de índices, de geração de índices a partir de tesauros, dentre outros. A análise destas ferramentas leva-nos a crer que ainda existe uma carência de junção destas soluções em um ambiente integrado, pois em geral elas pressupõem automação de atividades de forma isolada. Não geram tendências ou compartilhamento com outras ferramentas e acabam por perder o valor atribuído pelos profissionais de informação. Os sistemas que automatizam a construção de tesauros representam muito mais um esforço de gerenciamento pelos usuários de informação do que um auxílio à geração do próprio tesauro e

exigem ainda do usuário uma manutenção manual, sujeita à morosidade do próprio processo. Aos profissionais de informação, que na realidade são os usuários gerentes desses processos, interessa uma solução que crie conexão entre esforços e possibilite maior agilidade dos processos de indexação. Uma solução comprometida com a performance do processo sem esquecer a qualidade do produto gerado.

Assim, o problema existente pode ser resumido na falta de um instrumento integrado para a geração e manutenção de tesauros, com funções semi-automatizadas, capaz de selecionar e sugerir termos e relacionamentos entre estes, indicando aos usuários gestores do processo a pertinência dos mesmos frente a uma coleção de documentos.

Em função dessa ausência, este trabalho propõe a criação de um ambiente para geração e manutenção semi-automática de tesauros que traga integração entre ferramentas já existentes, com acréscimos ou uso diferenciado para as mesmas, e que facilite os processos de geração e manutenção de tesauros retirando dos usuários gestores tarefas possíveis de serem realizadas de forma automatizada. Não se pretende automatizar o processo de criação de tesauros mas acrescentar a ele facilidades, de forma que o indexador utilize seu potencial crítico nas ações intelectuais do processo apoiado pela tecnologia.

1.2 Objetivos

Este trabalho teve como objetivo geral propor uma metodologia e um ambiente para a geração e manutenção de tesauros de forma semi-automatizada, com utilização da linguagem natural, tecnologias da Ciência da Computação e

fundamentos teóricos da Ciência da Informação, mais especificamente, através dos conceitos ordenadores da garantia literária, da garantia de uso e da garantia estrutural e da garantia da própria estrutura do texto.

Como objetivos específicos pretendeu-se:

- Elaborar uma metodologia para geração e manutenção semi-automática de tesauros a partir da linguagem natural encontrada em textos científicos;
- Implementar ferramentas para verificação da atualidade e o potencial representativo de um tesauro;
- Avaliar e testar algumas tecnologias existentes para auxílio aos processos de geração e manutenção de tesauros.

1.3 O método

O homem sempre fez de sua observação e da própria experiência vivida a base para a construção de seu saber. O conhecimento do fogo e de como produzi-lo é um exemplo. Com certeza a própria natureza já teria proporcionado a visão de queimadas devido às altas temperaturas e o homem já conhecia o conforto que o fogo podia trazer à sua vida e por isso o mantinha (LAVILLE, 1999). Mas, ao aprender que friccionar pedras ou galhos secos produziria fogo, mudou a relação entre o homem e esse fenômeno. Já era possível controlá-lo e produzi-lo quando assim lhe conviesse. Mesmo que a produção do fogo tenha começado como um pequeno incidente, interessa que o fato de dominar sua produção modificou a forma de lidar com esse objeto.

Neste ponto deveria estar todo e qualquer objetivo do trabalho científico: conhecer o funcionamento das coisas para poder interferir em acontecimento e dele tirar proveitos para melhorar a vida; mas, nem sempre trabalhos científicos significam melhoria na vida da coletividade.

Algumas vezes o homem recorreu a mitos e crenças para a explicação dos fenômenos e situações que vivenciava. Em outras, fez de sua própria intuição o pilar para explicação da realidade, agregada à tradição do saber que atravessava gerações e permanecia porque assim era no passado e sempre assim o seria. Mas, já bem cedo se questionava a fragilidade desse saber intuitivo, do saber oriundo do senso comum ou da tradição. A racionalidade era almejada e os filósofos cumpriram seu papel desde a antiguidade grega. Na Idade Média, à reflexão filosófica juntaram-se dogmas religiosos e a Teologia ganhou o espaço da Filosofia. O período do Renascimento foi marcante pela renovação nas artes e nas letras, mas não no domínio do saber científico. No século XVII floresceram idéias de observação empírica do real, para posterior interpretação dos fatos sob a luz das ciências matemáticas (LAVILLE, 1999). Começou então a definir-se a ciência experimental e o saber passou a ser parte não somente da observação, mas igualmente da experimentação e da mensuração.

Durante os séculos XVIII e XIX o método experimental desenvolveu-se por meio de múltiplas aplicações e foram muitas as descobertas que melhoraram a vida do homem. Laville (1999) afirma que o homem do século XIX foi o primeiro a morrer em um mundo bem diferente do que viu ao nascer. A ciência proporcionou progresso e qualidade de vida, muito embora acompanhados de sérios problemas no campo social. O século XIX assistiu também ao

desenvolvimento das ciências humanas através das especulações dos filósofos em suas percepções das sociedades e das formas de governo. Surgiram idéias de direitos e igualdades, liberdades sociais e econômicas.

O século XX introduziu mudanças nos padrões culturais com os quais percebemos, sentimos e agimos no e sobre o mundo, principalmente através das idéias de Einstein publicadas em 1905 sobre a teoria da relatividade e da radiação eletromagnética. A exploração do mundo atômico trouxe aos cientistas uma nova realidade alterando mais uma vez a visão de mundo e a forma de pensar. Os novos conhecimentos da física exigiam mudanças profundas nos conceitos de espaço, tempo, matéria, objeto e causa e efeito, alterando o modo de vivenciar o mundo. O universo passou a ser visto como um todo dinâmico, indivisível, cujas partes estão essencialmente inter-relacionadas (FREIRE, 2002).

Estes elementos influenciaram as Ciências Sociais. Nelas, o foco é o homem em suas relações com os outros homens e a realidade. No Brasil, o CNPq organiza a área de conhecimento Ciências Sociais através das áreas de direito, administração, economia, arquitetura, planejamento urbano, demografia, ciência da informação, museologia, comunicação, serviço social, economia doméstica, desenho industrial e turismo. Cada uma delas vê o homem sob determinado foco, de acordo com seu objeto de estudo, mas todas interessam-se primeiramente pelo homem. A Ciência da Informação não poderia ser diferente: interessa-se pela produção, coleta, tratamento, organização e disseminação da informação a partir do homem. É através dele e de suas relações sociais que a informação é produzida, percebida e legitimada para depois ser tratada, armazenada e difundida.

A Ciência da Informação estabeleceu-se como área do conhecimento em meados do século XX, quando então um grande volume de informações advindas do período de guerra era disponibilizada à sociedade, através de documentos e da produção científica, requerendo uma organização. A Ciência da informação nasceu no *"entremeio contraditório entre disciplinas sociais e tecnológicas e no espaço deixado por recortes já instituídos pela biblioteconomia e demais ciências sociais"* (MOSTAFA, 1996). Mas herdou das ciências sociais sua metodologia de pesquisa, pois seu norte continua sendo o homem em suas relações, agora, com o objeto informacional.

Em sentido amplo, o método adotado em uma pesquisa diz respeito à escolha de procedimentos sistemáticos para descrever e explicar os fenômenos observados (RICHARDSON et al., 1999). Dentre outros, o método experimental é aplicado fortemente pela psicologia social, para, principalmente, interpretar reações individuais em função das interações entre indivíduos (JESUÍNO, 1986). A Ciência da Informação, interdisciplinar em sua natureza, também acolhe instrumentos de outras áreas do conhecimento em suas pesquisas. A proximidade com a Ciência da Computação dá-se através dos recursos que esta ciência oferece para armazenamento e manipulação de dados, que, aos olhos da Ciência da Informação, poderão se transformar em informação através do sujeito, justamente nas relações que ela escolhe como objeto de estudo. A Ciência da Computação, pertencente ao grupo das Ciências Exatas e da Terra na classificação do CNPq, tem no experimento sua forma de busca para a solução de problemas, apoiada por modelos construídos na própria ciência ou em base matemática.

Neste projeto, optou-se pelo experimento como forma de proposta e análise de uma solução para o problema levantado: a montagem de um ambiente semi-automatizado para criação e manutenção de tesauros como base a esses processos, através de tecnologias da computação.

Partiu-se da hipótese de que a reunião das palavras-chaves existentes em artigos científicos poderia compor um conjunto de termos para a formação de um tesauro que atendesse às necessidades de indexação de artigos da coleção da qual tivessem sido retiradas e de outras coleções de artigos do mesmo tema ou da mesma área de conhecimento. Este conjunto de palavras deveria ser aplicado em verificações de freqüência de sua ocorrência nos títulos, nos resumos, no corpo dos textos e nas bibliografias dos artigos. As estatísticas proporcionadas pelo método deveriam auxiliar ao gestor do ambiente criado na composição do tesauro.

Observa-se que a estatística é utilizada em grande escala na pesquisa em Ciências Sociais, principalmente na área da economia. Os métodos quantitativos representam, conforme Richardson et al. (1999) a intenção de garantir a precisão dos resultados, evitando distorções de análise e interpretação. O trabalho de Salton (1981), com levantamentos estatísticos da freqüência de termos nos textos é um bom exemplo disto.

O ambiente semi-automatizado veio como verificação desta hipótese.

2 Revisão de Literatura

2.1 Pelos caminhos da indexação

As bibliotecas nasceram na Antiguidade. A Biblioteca de Alexandria, considerada a primeira biblioteca do mundo, foi fundada entre 331 e 330 a.C.. A palavra biblioteca está relacionada ao conceito de coleção pública ou privada de livros e documentos congêneres, organizados para o estudo, leitura e consulta. Pode significar também o prédio, ou até mesmo o móvel onde se encontra o acervo. O vocábulo origina-se do grego *bibliothéke*, que deu origem a *bibliotheca* no latim (FERREIRA, 1989).

Desde a Antiguidade, os bibliotecários se preocuparam com a organização dos acervos. Começaram por gerar listas como forma de facilitar o encontro de documentos. Há registros de bibliotecas na era pré-cristã, que possuíam um catálogo de obras gravado em pedras. No século XIV, os monges criaram listas completas das obras dos mosteiros, indicando o lugar onde se encontravam, assim como compilavam listas de várias bibliotecas, tendo em vista que visitavam diversos mosteiros. Surgiram os guias para livros isolados, os resumos, os cabeçalhos de capítulos, os textos em margens, os índices do próprio livro, etc. Todos eles significaram pontos de avanço e alcance de uma indexação de acervos, na tentativa de sempre facilitar o seu manuseio (COLLISON, 1972).

Embora já existisse anteriormente como atividade, a definição da documentação e de sua função são recentes. Mesmo atrelado ao processo de armazenamento dos documentos, o conceito de documentação veio junto à

idéia de exploração da informação armazenada. Sua história inicia-se no século XIX, com Paul Otlet (1863-1944) que cria, em 1892, com Henri La Fontaine o *Office International de Bibliographie*, em Bruxelas. Ambos trabalharam na tradução para o francês da quinta edição da Classificação Decimal de Dewey (CDD) quando propuseram inovações permitindo uma classificação não apenas enumerativa, mas capaz de levar à construção de números compostos indicando assuntos inter-relacionados. Em 1895 eles transformaram o órgão criado no *Institut International de Bibliographie*, empreenderam a redação de um repertório, em fichas, de acordo com uma classificação de autores e uma classificação sistemática. Em 1910 celebrou-se o primeiro Congresso Mundial de Bibliografia.

Em 1912, o microfilme apareceu, permitindo o armazenamento da informação em formato reduzido. No período seguinte à primeira guerra mundial (1914-1918) privilegiou-se a importância da utilização dos serviços das bibliotecas. Nos anos 30 foram efetuados os primeiros trabalhos no domínio das técnicas documentais. As revistas e os periódicos começaram a ter relevância. Em 1931 o termo "documentação" passou a ser empregado e o *Institut International de Bibliographie* recebeu o nome de *Institut International de Documentation*. No ano seguinte criou-se na França a *Union Française des Organismes de Documentation*, primeiro organismo francês de documentação. O ano 1934 representa uma data importante na história da documentação, com a publicação do Tratado de documentação de Paul Otlet que constitui uma obra fundamental da área.

No final dos anos 30, a documentação afirmou-se, organizando-se no plano mundial. Destacam-se dois eventos: o primeiro Congresso Mundial de

Documentação (Paris, 1937) e a criação, em Haia, da Federação Internacional de Documentação (FID). Paralelamente desenvolvem-se as técnicas documentais com o uso das máquinas de cartões perfurados, utilizadas principalmente nos Estados Unidos a partir dos anos 40. Também foi lá e nesta mesma época que começaram a ser utilizados os seletores fotográficos.

Após a segunda guerra mundial, o desenvolvimento dos processos e das técnicas documentais acelerou-se em diversos domínios. Em 1949, a Unesco patrocinou a Conferência Internacional sobre a Análise dos Documentos Científicos. Verifica-se um aumento considerável de publicações sobre o tema na literatura. Em 1950 criou-se na França a primeira instituição de ensino da documentação: o *Instituit National d'es Techniques de la Documentation* (INTD) ligado ao *Conservatoire National d'es Arts et Métiers* (CNAM).

Por volta de 1952, Mooers publica os seus trabalhos, com a descrição do sistema Zatocoding, atribuindo-se a ele a expressão "*information retrieval*" (CHAUMIER, 1971). A seguir, surgiram novas propostas para registros: o sistema de Mortimer Taube ou sistema Uniterm (1953), o sistema do índice permutado de Citron, Hart e Ohlman e o sistema de divulgação seletiva de informação de Luhn, estes dois últimos em 1958. As reuniões e as conferências internacionais tornaram-se mais numerosas. Os anos 1960 são caracterizados pelo fenômeno da explosão documental que se traduz por um aumento das publicações científicas e técnicas e uma multiplicação dos centros de documentação ou de bibliotecas. Na França existiam 2352 centros de documentação em 1962 (CHAUMIER, 1971).

Nos anos 1970 os computadores cresceram em potencialidade, permitindo o tratamento em tempo real; começam os trabalhos em redes. No fim deste período histórico cabe uma redefinição do que seja documento: o documento não é mais apenas o livro, mas também o artigo de periódico, a ata de um congresso, um relatório de investigação, sem esquecer os novos suportes de informação: os filmes, as microfichas, os arquivos magnéticos. Quanto à documentação, trata-se de uma disciplina científica que recorre à lingüística, à matemática e à informática. A documentação "*tornou-se verdadeiramente o tratamento da informação não numérica em todas as suas formas*" (CHAUMIER, 1971).

2.2 Linguagens

A heterogeneidade pode ser observada na própria realidade, quer na natureza das coisas, quer em nossa própria natureza. Somos indivíduos de uma mesma sociedade, utilizamos os mesmos serviços, freqüentamos lugares comuns, e somos diferentes. Visto de um ângulo maior, nos constituímos enquanto grupos de acordo com nossos interesses, raças, culturas, afinidades, conhecimento, etc. Como animais que somos, nos distinguimos destes não pela capacidade de aprender, mas pela qualidade de sabermos transmitir o que aprendemos. O homem é um animal que conquista e constrói não só para si, mas para toda sua raça. Algumas espécies buscam na caça sua sobrevivência ou, no máximo, a sobrevivência de seu grupo ou família. O homem busca formas para garantir não só seu sustento ou sua vida, mas o de toda sua raça. O homem aprende para a coletividade, e na coletividade constrói seu conhecimento.

Por ser um ser social e viver em coletividade, existe no homem o grande desejo de relacionamento com seus iguais. Chega a ser uma necessidade humana não estarmos isolados uns dos outros. No âmbito social, formamos grupos orientados por interesses e desejos comuns, pela distribuição geográfica, por nossa raça e, principalmente, por nossa cultura. Mesmo assim, insistimos a todo instante em nos comunicarmos uns com os outros, sendo vital esse relacionamento. Para estabelecer tais relações, Curras (1995) aponta dois pontos importantes: primeiro, a linguagem, fundamental para que o processo exista; segundo, a informação, conseqüência de tudo isto. Para essa autora, a linguagem não é algo estático, possui seu movimento contínuo e evolutivo, assim como nos encontramos em contínuo movimento e evolução. É preciso conhecer esta mudança nos idiomas para que nossa comunicação siga seu caminho natural. Daí justifica-se a importância atual da terminologia, da tradução e da normalização.

Por linguagem entende-se a manifestação dos seres vivos, empregada em suas relações de convivência existencial. É um instrumento utilizado para a comunicação entre os homens, constituído de um conjunto de símbolos que estão combinados de forma sistemática, para o armazenamento e troca de informações. Neste trabalho, a palavra linguagem estará ligada ao contexto humano, ao artifício que utiliza o homem para se comunicar com outros homens. Antes de tudo é um instrumento de mediação e pode ser vista como *"um conjunto de símbolos convencionais, portadores de informação, cujo principal fim refere-se à comunicação entre os seres vivos"* (CURRAS, 1995).

As linguagens exercem influência nos sistemas de recuperação de informação. Estes sistemas possuem, cada qual a seu modo, uma forma de busca

específica que os caracteriza e os distingue em seu universo. Uma estratégia de busca é uma técnica ou um conjunto de regras que são aplicadas para tornar possível o encontro entre o desejo de um usuário, formulado por uma requisição ao sistema, e a informação nele armazenada. De forma geral, as máquinas de busca na Web utilizam-se de linguagem natural para os processos de recuperação de informações. Para Lopes (2002), a Linguagem Natural pode ser conceituada como *"sinônimo de discurso comum"*, ou seja, é a linguagem que uma comunidade utiliza em seu dia a dia, habitualmente na fala e na escrita. É também conhecida como vocabulário livre.

Chaumier (1971) classifica a linguagem natural em três grandes campos de estudo: o estudo dos sons ou a fonologia pelo qual se interessa o profissional de documentação na tradução automática de simbolização fonética; o estudo das palavras ou a lexicologia, de primordial importância para a área de documentação, encontra-se na base de qualquer linguagem de documentação; o estudo da sintaxe, que intervém quando uma linguagem de documentação tem um caráter elaborado ou complexo.

Lancaster (1993) afirma que a linguagem natural é, também, a linguagem utilizada no discurso técnico-científico. Os sistemas de recuperação utilizam-se dela nos processos prévios de indexação das palavras encontradas nos textos que servirão para localização dos mesmos, no instante das requisições dos usuários.

Á Ciência da Informação, como a outras áreas do conhecimento, importa como as linguagens especializadas são utilizadas por grupos de usuários, com fins definidos. O tema linguagens é aprofundado por diversos cientistas, que se

Interessam, sobretudo, pelas linguagens de indexação, também conhecidas como linguagens de documentação ou linguagens documentárias, utilizadas nos serviços de armazenamento, indexação e recuperação da informação, processos básicos nessa ciência. Neste trabalho, os termos linguagem controlada, linguagem de documentação e linguagem documentária são tomados como sinônimos de linguagem de indexação.

Para Svenonius (2000) um documento é visto como uma incorporação particular em espaço e tempo da informação. A linguagem utilizada pelo documento, a linguagem natural, descreve e fornece o acesso a esta incorporação. É distinta de uma linguagem que ela classifica como linguagem do trabalho, ou seja, a linguagem com a qual se documenta atributos específicos do documento e que provê acesso ao documento – o que é semelhante ao que se denomina linguagem de documentação. O que distingue essas linguagens é o fato da primeira enfocar o pensamento e a segunda a sua manifestação. A primeira é fruto do pensamento dos autores de cuja elaboração nasceu o documento. A segunda é a forma como esse pensamento passa a ser representado, ou seja, sua manifestação. Assim, a linguagem do documento pode também ser vista como a língua que descreve e fornece o acesso às manifestações dos trabalhos ou do conteúdo da informação.

Consideram-se linguagens de indexação os instrumentos de representação de informação para a indexação, o armazenamento e a recuperação de documentos. As mais conhecidas são os tesauros e os sistemas de classificação bibliográfica. São linguagens artificiais que foram construídas e elaboradas. Não resultam de um processo evolutivo e necessitam de regras explícitas para seu uso. Um vocabulário controlado, também chamado de

linguagem controlada, é uma linguagem documentária, um conjunto de termos organizados sob determinada forma - hierárquica e/ou alfabética - cujo objetivo é possibilitar a recuperação de informações temáticas, reduzindo a diversidade da terminologia.

Para Lancaster (1993) e para Lancaster e Warner (1993) um vocabulário controlado é essencialmente uma lista de termos autorizados, com estrutura semântica, que se destina, principalmente, ao controle dos sinônimos, à diferenciação de homógrafos e à reunião ou ligação dos termos. Para o autor, os termos utilizados no processo de indexação serão extraídos de algum tipo de vocabulário controlado, como um tesauro, ou poderão ser utilizados termos livres e extraídos do próprio documento.

A classificação das linguagens de documentação para Chaumier (1978) dá-se de acordo com os seus tipos e com o seu desenvolvimento histórico linguagens. Podem ser classificadas nas seguintes categorias: as linguagens de classificação, que abrangem a classificação decimal e a classificação por facetas; as linguagens controladas, que envolvem as listas e os vocabulários e as linguagens de descrição de dados.

As listas de vocabulários são utilizadas em bibliotecas e constituem sistemas pré-coordenados de indexação. Em geral são apresentadas em estruturas de verbetes amplos e específicos, apresentando, quando muito, as relações de equivalência. Este tipo de recurso ainda é utilizado para certos índices.

Para Chaumier (1978), a ausência de sintaxe em linguagens de documentação era uma característica comum, o que ocasionava dificuldades no seu emprego, trazendo como conseqüência uma redução no desempenho dos sistemas

documentais. Alguns pesquisadores tentam propor linguagens com uma sintaxe – por menor que seja – que permita sanar as principais ambigüidades existentes nas linguagens sem sintaxe. O autor chama a atenção para os trabalhos de R. Pàges com a sua linguagem conhecida como de *"análise não codificada"* e de J. C. Gardin e sua equipe, com a célebre *Systematic Organized Language* (SYNTOL).

2.3 Trabalhando com as linguagens de indexação

Todo esforço realizado nos Sistemas de Recuperação de Informação possui um só objetivo: possibilitar ao usuário o acesso à informação. São diversos os instrumentos utilizados para a consecução desse objetivo e envolvem sempre diversas áreas do saber. Mas, para que a informação possa ser disponibilizada é necessário primeiramente que ela esteja organizada para tal. A classificação é um instrumento de organização da informação.

Os esquemas de classificação existem desde a Antiguidade, em conotação distinta da utilizada na Ciência da Informação. No decorrer da história, os lógicos, os filósofos e os lexicógrafos utilizaram-se da classificação para compreender e analisar o conhecimento, interpretando o significado da classificação conforme era conveniente às suas idéias. Aristóteles (382-322 a.C.) percebia a classificação como um "*exercício mental*" e em sua obra Organon apresentou suas grandes categorias para classificação (KAULA, 2005): Gênero, Espécie, Diferença, Propriedade e Acidente. Estas desdobraram-se em 10 categorias (Substância, Qualidade, Quantidade, Relação, Lugar, Tempo, Situação, Posse, Ação e Sofrimento ou passividade)

que foram usadas pelos aristotélicos e outros para qualificar os conceitos das diversas áreas do conhecimento.

Para Aristóteles, tudo o que existia no mundo físico poderia ser expresso em termos de um conjunto de características que individualizam e tornam real um objeto frente aos demais. Através desse conceito de realidade a partir do conjunto de atributos que possuíam as coisas, derivou-se a ciência da classificação de Aristóteles em gêneros, espécies e sub-espécies. E foi justamente destes princípios que surgiu a ciência que durante séculos utilizou-se da observação de ocorrências na realidade e, com a devida generalização indutiva, criou classificações gerais de fenômenos físicos de acordo com seus atributos e modos de funcionamento (SHERA, 2005).

Mas esta classificação aristotélica pecou ao desconsiderar a interconectividade existente entre os objetos do mundo real. Nada no universo é completo em si mesmo, sem necessidade de referência a qualquer outra coisa substancial e nem tampouco de ser o universo fragmentado em milhares de substâncias desconectadas. Apesar de tudo, essa foi a lógica com a qual se construiu a classificação taxonômica, que durante séculos orientou o trabalho de classificação de documentos em uma ordem rígida de hierarquias (SHERA, 2005).

Apenas no final do século XVIII o termo ‘classificação’ foi relacionado com a apresentação de um plano para a classificação das ciências e dos livros (DAHLBERG, 1972b).

No século XIX, especialmente, a elaboração de tais planos tornou-se um hobby para cada filósofo, bem como para alguns cientistas. Ampère, no prefácio de sua classificação de 1834-1843, escreveu:

> *"Já há algum tempo eu compreendera que, ao tentar determinar as características distintivas para a definição e classificação das ciências, é necessário considerar não só a natureza dos objetos à qual eles se relacionam, também os pontos de vista sob os quais esses objetos podem ser considerados..."* (AMPÉRE[1] *apud* DAHLBERG, 1972b).

Na biblioteconomia, foi Henry Evelyn Bliss (1870-1955) o primeiro a considerar esses pontos. Bliss preocupou-se principalmente com os fundamentos filosóficos de classificação. Em seu sistema final de classificação, publicado primeiramente em 1935, depois revisto e ampliado nas edições de 1940 e 1953, Bliss apresentou os diferentes aspectos de cada área também de forma diagramática, geralmente em nível hierárquico, embora algumas vezes em arranjos bidimensionais. A grande contribuição de sua obra foi o contato com os princípios filosóficos da classificação, entre eles, os fundamentos conceituais da formação, divisão e partição de classes. Dahlberg (1979b) aponta que a maior contribuição de Bliss foi ter proporcionado ao indiano Ranganathan a mais fértil das inspirações.

2.3.1 Teoria da Classificação facetada

A teoria da Classificação Facetada foi elaborada por Shiyali Ramamrita Ranganathan nos anos 30, a partir da *Colon Classification*, tabela de classificação, elaborada e adotada na Biblioteca da Universidade de Madras,

[1] AMPÈRE, A. M. Essai sur la philosophie des sciences, or exposition analytique d'une classification naturelle de toutes les connaissances humaines. Partie 1-2. Paris, 1834-1843.

Índia. Até então, as tabelas empregadas nos processos de documentação não apresentavam as bases teóricas utilizadas em sua elaboração. Esta teoria recebeu este nome por adotar uma técnica de fragmentar um assunto complexo em diversos aspectos ou partes denominadas facetas (TRISTÃO; FACHIN; ALARCON, 2004).

Ranganathan nasceu na Índia em 1892, onde morreu em 1972. Tinha o título de mestre em Matemática e ensinava na Universidade de Madras. Em 1924 foi designado como bibliotecário da universidade, mantendo-se na função por cerca de vinte anos. Já como bibliotecário, ele viajou para a Inglaterra onde estudou na *London's School of Librarianship.* Mostrava-se descontente com a Classificação Decimal de Dewey muito embora esse fosse o sistema mais popular de classificação utilizado naquele país. Para ele o sistema continha deficiências para expressar a totalidade dos aspectos do assunto específico de um documento. Ranganathan tinha idéia de um sistema que permitisse a inclusão de assuntos ainda inexistentes ou não planejados. Suas idéias estão apresentadas em suas principais obras: *Prolegomena to Classification* (1937), *Philosophy of Book Classification* (1951) e a própria *Colon Classification* (1933). Em todas elas sente-se o traço da filosofia oriental e da matemática e é justamente esta integração entre o pensamento racional e o pensamento oriental que coloca a Teoria da Classificação Facetada em um espaço singular (CAMPOS, 2001).

Ranganathan levantou o problema da dinâmica do surgimento dos assuntos tratados nos documentos frente a estrutura classificatória utilizada na época. O fato em si não é novidade, já que outros autores haviam levantado a mesma questão. Para ele, a totalidade das idéias conservadas pelo ser humano se

apresentava em um contínuo dinâmico; era preciso uma nova teoria capaz de responder a essa necessidade (CAMPOS, 2001).

No contexto das atividades de classificação, Ranganathan observa três planos de trabalho: O Plano da Idéias, o Plano Verbal e o Plano Notacional. Cada um deles possui princípios normativos próprios e cumpre objetivos específicos. No Plano das Idéias é onde existe a formação do processo do pensar, onde originam-se as idéias. Para KAULA (2005) o criador das idéias passa por um processo de autocomunicação no interior da mente e surgem mais idéias. Para esse autor, o trabalho no plano das idéias pode ser tomado como análise do conceito.

> *"Uma idéia é um conceito que ao tomar forma concreta pode levar a alguma informação. A análise conceitual é uma tarefa difícil que tem que ser esgotada na concepção do esquema de classificação. Um conceito pode ser um isolado, um quase isolado ou um assunto e é a identificação de conceitos, sua posição no universo de assuntos, seu arranjo sistemático entre outros conceitos, etc., que faz do trabalho uma tarefa árdua"* (KAULA, 2005).

O Plano Verbal objetiva permitir que a linguagem seja uma mediadora para a comunicação das idéias e conceitos. Para tal, precisa ser livre de homonímia e sinonímia, principalmente por se tratar de uma linguagem classificatória que não é uma linguagem natural (CAMPOS, 2001; KAULA, 2005).

O Plano Notacional é o da representação, dos números que passam a representar os conceitos. É mais restrito, pois tem que trabalhar de acordo com o que foi produzido no plano das idéias. O Plano Notacional obedece firmemente o que foi decidido no Plano da Idéias, muito embora seja

necessário desenvolver versatilidade com o objetivo de complementar totalmente as descobertas no plano das idéias (KAULA, 2005). Encontra-se aí o princípio da hospitalidade, ou seja, a propriedade representativa da capacidade existente na estrutura classificatória em abrigar os novos conceitos oriundos do Plano das Idéias.

Campos (2001) afirma que Ranganathan estende o conceito de hospitalidade, que antes permitia apenas a criação de sub-classes nas classes já existentes, impedindo a criação de novas classes. Para a autora, o princípio aplicado com esta abrangência torna possível a introdução de mecanismos para que o esquema classificatório acompanhe a dinâmica do conhecimento. E é possível a ampliação da base notacional, a ampliação dos renques e a organização da estrutura classificatória em categorias fundamentais: Personalidade (*Personality*), Matéria (*Matter*), Energia (*Energy*) , Espaço (*Space*) e Tempo (*Time*) – PMEST. Assim, o elemento utilizado na classificação não é mais o assunto do documento, *"mas os conceitos que fazem parte deste assunto, organizados a partir das categorias dentro da área do conhecimento"* (CAMPOS, 2001).

A extensão do conceito de hospitalidade permitiu também a adoção do método analítico-sintético, demarcando a separação entre os momentos da elaboração de esquemas de classificação, da análise do documento e do uso desse esquema. A classificação facetada é um esquema analítico-sintético, pois envolve dois processos distintos: a análise do assunto em facetas e a síntese dos elementos que constituem o mesmo, sendo, portanto, aplicável a qualquer área do conhecimento (TRISTÃO; FACHIN; ALARCON, 2004).

Os elementos da estrutura classificatória de Ranganathan estão organizados a partir de suas características, em renques e cadeias, distribuídos em facetas e estas organizadas em categorias. As unidades classificatórias utilizadas são o assunto básico e a idéia isolada. O assunto básico pode ser visto como um assunto sem nenhuma idéia isolada como componente. É um corpo sistematizado de idéias de um campo especializado e representa as áreas mais abrangentes do conhecimento. São exemplos de assunto a Matemática, a Agricultura, etc. A idéia auxilia na compreensão do assunto. Por exemplo, Milho é uma idéia isolada que aplicada junto ao assunto Agricultura forma outro assunto: Agricultura do milho (CAMPOS, 2001).

Ranganathan utiliza os conceitos de características para comparar os elementos classificatórios, tendo em vista a formação de classes e, dentro delas, os renques e as cadeias. As características estão associadas a propriedades comuns dos objetos. Os renques derivam-se de uma única característica em algum passo de divisão de características, formando séries horizontais.

> *"O renque é constituído de conceitos subordinados a um mesmo conceito, ou seja, conceitos coordenados; são conceitos 'irmãos'. O renque é, portanto, uma série horizontal de conceitos. Por exemplo, 'água mineral' e 'água potável' têm o termo superordenado, 'bebida hídrica natural'; logo, são de mesmo nível de coordenação, constituindo-se num renque"* (BITI, 2005).

Uma cadeia é uma seqüência formada por classes e através do deslocamento entre cadeias é possível chegar ao ponto desejado. Essa seqüência é formada por séries verticais de conceitos onde cada qual possui uma característica a

mais ou a menos. Assim, os renques e as cadeias formam a organização da estrutura classificatória que é totalmente hierárquica. Ranganathan estabelece uma série de regras (cânones) para estabelecer a conduta uniforme na formação dos renques e cadeias, entre eles o cânone da Exaustividade, o da Exclusividade, o da Seqüência útil, o da Modulação, etc. (CAMPOS, 2001).

Uma faceta é um termo genérico utilizado para indicar algum assunto básico ou isolado, tendo também a função de formar renques, termos e números (RANGANATHAN, 1967). As facetas são de dois tipos: - a faceta básica que agrupa assuntos básicos (áreas do conhecimento) e a faceta isolada que agrupa conceitos isolados.

De uma forma mais genérica, Svenonius (2000) explica que o maior problema relacionado às facetas é a dificuldade de se definir as próprias facetas e conseqüentemente em se definir classes dentro delas. Para ela, estas dificuldades são causadas pelo fato de que a língua é algo muitas vezes sem regras. Existem os termos que não podem ser categorizados exclusivamente em apenas uma faceta e termos que resistem à categorização porque são muito ambíguos ou abstratos. Mais do que isso, definir facetas por critérios semânticos e sintáticos (como muitas vezes é feito) conduz a uma classificação transversal e inconsistente. Estas são as dificuldades fundamentais encontradas nas raízes escondidas de muitas classificações baseadas em lingüística.

O postulado das categorias é o princípio normativo para se organizar universos de assuntos. O primeiro passo é justamente mapear o universo, seja ele facetado ou enumerativo. Ranganathan postula que existirão, em todos os

universos de assunto, cinco idéias fundamentais: PMEST. A categoria tempo é definida com seu entendimento usual, ou seja, ao período a que está associado o assunto. A categoria Espaço também o é, exemplificando: as regiões geográficas da localização do assunto. A categoria Energia pode ser compreendida como uma ação de uma espécie que ocorre com respeito ao assunto. A categoria Matéria é a manifestação de materiais em geral, com suas propriedades ou o constituinte material de todas as espécies. A categoria Personalidade não possui definição. E nela estarão os que não forem definidos como espaço, energia ou matéria. A reunião de todas as categorias forma um sistema de conceitos de determinada área de assunto sendo que cada conceito desta categoria é também a manifestação da mesma (CAMPOS, 2001).

Lima (2004) aponta que a elaboração de uma classificação facetada parte do exame da literatura do assunto na identificação dos conceitos e termos, estabelecendo características e então, dentro delas, as facetas. Para a autora, *"a faceta é a coleção de termos que apresentam igual relacionamento com o assunto global*" (LIMA, 2004), agrupados por princípio básico de divisão. Cada assunto tem suas próprias facetas sendo que novos agrupamentos podem surgir dentro de uma faceta. Importa também, que os termos organizados em facetas não se sobreponham.

2.3.2 Teoria da Terminologia

Na esfera da informação, a terminologia é peça chave para os processos de sua produção, de seu armazenamento e de sua recuperação. O vocábulo *Terminologia* tem origem na junção do prefixo latim *terminu*, que significa *termo*, acrescido do sufixo *logo*, no sentido de tratado, de acordo. Nos

dicionários, encontramos sentidos que remetem ao geral ou específico: pode ser vista como o tratado dos termos específicos de uma arte ou ciência, como o conjunto de todos estes termos ou como o emprego de palavras peculiares a determinado escritor, a uma região, a uma religião, etc. (DICIONÁRIO UNIVERSAL DA LÍNGUA PORTUGUESA, 2005). Para Campos (2001), a esses significados deve-se incorporar a terminologia como um conjunto de princípios teóricos. A ciência possui sua própria terminologia, subdividida nas terminologias específicas de cada uma de suas áreas.

A autora lembra que o primeiro significado está diretamente ligado com a lexicologia especializada, que é *"o estudo científico do conjunto de termos de uma dada língua, em uma área especializada"*. O segundo significado remete ao campo dos dicionários técnicos, dos vocabulários e léxicos. E então, a compreensão do que é terminologia passa a ser uma apresentação ordenada de um certo grupo de conceitos. Ver a terminologia como um conjunto de princípios teóricos é considerá-la como uma área de estudo científico o que propicia ao termo o status de área de saber. A *"ciência da terminologia significa: ramo do saber, disciplina científica, uma ciência em si, ciência como tal"* (DROZD, 1981 *apud* CAMPOS, 2001).

O vocábulo terminologia não existia originalmente no latim. Foi introduzido inicialmente no alemão e posteriormente incorporado ao inglês e francês no início do século. Neste trabalho, ao empregar o termo terminologia estaremos nos referindo ao conjunto dos termos próprios de uma comunidade, de uma ciência, de uma arte, de uma área específica do conhecimento ou atividade humana.

Se nos debruçarmos no tempo à procura da compreensão histórica de quando o homem começou a classificar e categorizar o real, certamente nos defrontaremos com o período pré-socrático, quando a natureza inspirava os filósofos no entendimento da vida, de seu movimento e de nossa própria razão de ser. Mas, seria ir longe demais. As próprias divisões da história constituem categorias que criamos para melhor compreender e estudar os fatos. Cada área do conhecimento cria suas próprias divisões históricas, conforme favoreçam a explicação dos fatos. Neste trabalho, utiliza-se o caminho percorrido por Curras (1995) para demarcar os períodos históricos que favoreceram o entendimento da questão terminológica:

- Primeiramente, um período que a autora denomina de clássico, desde as civilizações da Antiguidade até o final do século XVII, constituindo milhares de anos, quando aconteceram fatos marcantes da história da humanidade. Até o século XVII, no início do que conhecemos como ciência moderna, havia idéias universais das ciências, da filosofia e de todos os ramos do conhecimento. As idéias eram universais e os enciclopedistas foram um marco desse período. Naquela época, os especialistas se dedicavam a áreas distintas do saber, mas possuíam *"consciência de sua especialidade como parte de um todo"*. Já existiam particularidades nas linguagens utilizadas, diferenciadas por seu emprego e nas áreas em que trabalhavam.

- A seguir vem o período de iniciação da terminologia moderna, abrangendo os séculos XVIII e XIX. Destacam-se os trabalhos de Lavoisier e Bertholler na nomenclatura e classificação para a química.

Os descobrimentos científicos e as invenções técnicas que surgiram, trouxeram com eles uma grande revolução lingüística e terminológica. As atividades no campo da normalização do trabalho também merecem ser ressaltadas, pois levaram a uma normalização da linguagem para realização de tradução correta para os vocábulos empregados.

- O terceiro período vai do início do século XX até os anos 30 e é chamado de período da consolidação da área de terminologia, pois foi nele que se criou uma consciência do surgimento dessa nova disciplina, independente por si mesma e relacionada intimamente com a lingüística, com a lógica, com as ciências naturais e as teorias classificatórias. Os cientistas foram os propulsores do fato, pela própria necessidade de utilização de termos específicos na comunicação com seu próprio grupo. Em 1927, em Washington, realizou-se a 2ª Conferência de Padronização Pan-Americana e uma das conclusões postulava que o espanhol, língua oficial dos países hispano-americanos, deveria ser estudado e levado em consideração nos trabalhos de Lingüística e Tradução. Deveriam também evitar a diversidade de palavras que se utilizavam em diferentes países para as mesmas coisas e os mesmos conceitos. Em 1938, o IEC[2] publicou o *International Eletrotechnical Vocabulary*, outro marco representativo do período.

[2]O *International Electrotechnical Commission* (IEC) é uma organização direcionada para o estabelecimento de padrões relacionados às tecnologias da elétrica, da eletrônica e outras áreas relacionadas. Muitos de seus padrões são desenvolvidos juntamente com o *International Organization for Standardization* - ISO. Foi fundado em 1906 e hoje possui mais de 60 países participantes. Originalmente estava localizado em Londres, mas sua atual sede está em Genebra desde 1948. Este instituto distribui padrões para unidades de medida, particularmente o *gauss*, *hertz*, e *weber*. É também dele o Sistema de Giorgi, que em última instância tornou-se padrão para *International System of Units* – (SI). Em 1938, publicou um vocabulário de tradução internacional para unificar a terminologia elétrica, trabalho que se tornou contínuo e permanece de suma importância para as indústrias elétricas e eletrônicas.

- O período da teorização abrange de 1930 até 1970, quando foram instituídos os princípios teóricos da terminologia moderna. O fato mais marcante é a tese de doutoramento de Eugen Wuster sobre a normalização da linguagem na técnica, principalmente em eletrotécnica, considerado como um tratado da Teoria Geral da Terminologia. A originalidade de seu trabalho consiste em ter aplicado, como se supõe, as idéias de F. Saussure às linguagens científicas especializadas, "colocando a ciência em contato com a lingüística e a filosofia". Desse período data a criação da Escola de Viena e o surgimento de escolas em Praga e por toda a França. A Escola Húngara, também dessa época, firmou-se nos problemas relativos à tradução. Em 1942, A. M. Terpagoreve, terminólogo e presidente do Comitê de Terminologia Técnica e Científica (KNTT), expôs a necessidade de se iniciar os estudos para o estabelecimento de base teóricas para o Sistema de Conceitos, fato que direcionou os trabalhos da escola Soviética (CAMPOS, 2001). Em 1954, criou-se a União Latina, com o objetivo de defender, conservar e difundir os idiomas de raízes latinas: o espanhol, o português, o francês, o italiano e o romeno (CURRAS, 1995).

- Do início dos anos 70 até nossos dias, diversos trabalhos têm sido produzidos no tema e o alcance a que chegamos torna fácil denominar esse tempo como "período de evolução acelerada". Os fundamentos teóricos para a terminologia continuaram a ser estudados na Escola de Viena. Helmut Felber deu continuidade às idéias de Wuster e Christian Galinski foi seu sucessor. Nos últimos anos, o campo da terminologia foi-se expandindo, chegando aos sistemas classificatórios. Em 1971

deu-se a criação do centro Infoterm, relacionado com o programa Unisist e subordinado à Unesco. Trata-se de um centro internacional de informação sobre tecnologia, com o objetivo de apoiar e coordenar a cooperação internacional neste campo; influenciou a criação de centros nacionais de terminologia em diversos países como França, Alemanha, Espanha, Argentina, México e Canadá. No Brasil, a Embratel tornou-se um ícone entre as organizações que se interessam por assuntos terminológicos. Outro marco importante foi o Congresso sobre Terminologia e Engenharia do Conhecimento, realizado em 1987 em Trier, Alemanha, com trabalhos marcantes sobre terminologia em Informática e engenharia das línguas.

2.3.3 Teoria dos Conceitos

Um tesauro é um vocabulário de termos de determinada área do conhecimento, relacionados genérica e semanticamente. É uma forma de representação do conhecimento de determinada área do conhecimento. A teoria dos conceitos facilita a compreensão dos tesauros. Felber[3] (1984) apud Motta (2005) coloca que as primeiras investigações sobre a natureza dos conceitos foram feitas nas antigas escolas filosóficas gregas, sendo as idéias de Platão na obra Phaidron (teoria das idéias) consideradas o início dessa teoria. Mas, foi Aristóteles, na obra Organon, quem lançou os fundamentos da lógica que se tornaram leis básicas do conceito: características, raciocínio, inferência e definição, designadas por ele como *Analityk* (MOTTA, 2005).

[3] FELBER, Helmut. Terminology manual. Paris, Unesco/Infoterm, 1984.

Na década de 70, Ingrid Dahlberg incorporou ao campo da documentação os princípios terminológicos através de sua Teoria do Conceito, cujo princípio básico toma o conceito como uma representação do conhecimento. A teoria por ela introduzida possibilitou à Ciência da Informação uma compreensão maior do que se entendia por conceito e sua aplicabilidade na representação e recuperação de informação. Para Motta (2005) esta teoria destinava-se a servir de fundamento para análises conceituais de toda e qualquer iniciativa de estudo e padronização de termos. De maneira geral, para cada conceito temos um referente, sobre o qual podem ser realizadas afirmações passíveis de verificação. Todas essas afirmações podem ser sumarizadas por um termo, que passa a representar o conceito em qualquer processo de comunicação.

Tudo começa pela compreensão da própria linguagem natural, que permitiu ao homem explicar os objetos que encontrava na realidade através de símbolos ou palavras, possibilitando também a tradução de seu próprio pensamento. Assim, o conhecimento do homem fixou-se através dos elementos da linguagem. Para Dahlberg (1978) a linguagem possibilitou que o homem elaborasse enunciados sobre esses objetos, que podem ser considerados como individuais – aqueles pensados como únicos e distintos dos demais e cuja caracterização requer presença das formas do tempo e espaço - e os objetos gerais que, de certa forma, prescindem das formas de tempo e espaço. Aos objetos individuais estão associados o que ela chamou de conceitos individuais e aos gerais correspondem os conceitos gerais. A formação dos conceitos é a reunião e compilação de enunciados considerados verdadeiros a respeito de determinado objeto.

As características de um conceito traduzem seus atributos. Podem ser divididas em essenciais (necessárias) e acidentais (adicionais ou possíveis). As essenciais dividem-se ainda em características constitutivas de essência e características consecutivas da essência. As acidentais por sua vez dividem-se em características acidentais gerais e acidentais individualizantes. É através dessas características que se dá a análise do conceito. Em alguns casos as características podem estar hierarquizadas. Em outros, uma característica é apresentada de forma tão geral que pode ser considerada uma categoria. As características de um conceito possuem as funções de ordenação classificatória e dos respectivos índices, de definição do próprio conceito e de formação dos nomes dos conceitos. Conceito é então definido como

> *"... sínteses rotuladas de enunciados verdadeiros sobre objetos do pensamento: esses enunciados - asserções - levam ao reconhecimento ou à separação das características dos conceitos que também podem ser consideradas como os elementos dos conceitos"* (DAHLBERG, 1979b).

As relações entre conceitos podem ser definidas pela presença de características comuns em conceitos diferentes. Dahlberg (1978) mostra que estas relações podem ser: lógicas, hierárquicas, de oposição e funcionais. Esta última corresponde à inovação da teoria de Dahlberg, além da redefinição de conceito. De forma distinta das demais que pertencem a um corpo de conceitos estático, relacionando-se a objetos e propriedades, as relações funcionais desencadeiam-se por um processo ou uma atividade (MOTTA, 2005).

Os princípios descritos acima foram aplicados por Dahlberg ao reconhecimento, à construção e à utilização de sistemas de classificação. Hoje,

a teoria da classificação abrange o reconhecimento do conceito como elemento material dos sistemas de classificação e consideram a aplicação de uma teoria analítica de conceitos para a representação do conhecimento ou da informação (DAHLBERG, 1979a).

A própria autora indica as conseqüências de sua abordagem teórica da classificação: a avaliação dos sistemas de classificação existentes levando os mesmos a uma objetividade maior do que anteriormente; a construção de novos sistemas com agrupamentos ou arranjos previsíveis e a formalização de enunciados sobre o conteúdo de documentos que podem ser pesquisados com consistência, quer manualmente, quer por computador, a partir de estruturas de sentenças pré-determináveis (DAHLBERG, 1979a).

2.4 Tesauros

A palavra tesauro tem origem do latim *thesaurus*, que significa tesouro. Foi empregada como título no dicionário analógico de Peter Mark Roget, "*Thesaurus of English words and phrases*", publicado em Londres pela primeira vez em 1852. O autor era secretário da Royal Society e objetivava facilitar sua atividade literária. Trabalhou nesse projeto durante 50 anos. Em seu dicionário as palavras foram agrupadas em ordem distinta da alfabética. Priorizaram-se as idéias que exprimiam e esta foi a ordem escolhida. A busca por palavras dava-se sempre por aquilo que elas podiam expressar, com seu significado (GOMES, 1990).

A formação de um tesauro dá-se por palavras cuidadosamente escolhidas, palavras que possuam significado para uma determinada área. Roget chamou

sua obra de *thesaurus*, com o significado de vocabulário, dicionário. Alguns dicionários traduzem a palavra por enciclopédia ou tesouro. A originalidade deste trabalho foi que ele associou significação tão grande ao vocábulo que o mesmo permaneceu, para a área de documentação, associado à forma de organização do vocabulário para os processos de indexação e recuperação (CAMPOS, 2001; CHAUMIER, 1978; GOMES, 1990; MARQUES DE JESUS, 2002).

Motta (2005) citando Mikhailov et al.[4] (1973) afirma que o termo tesauro já havia sido utilizado anteriormente, tendo sido empregado por Brunetto Latini (1220-1294) para designar uma enciclopédia sistematizada que ele chamou de 'Os livros do tesouro'. Em 1532, Thierry e R. Etienne publicaram o livro 'Dicionário ou tesouro do idioma latino', um dicionário da língua latina em arranjo alfabético. Quarenta anos depois, A. Etienne publicou outro dicionário empregando o termo: "*Thesaurus linguage Graecae*", iniciado por seu pai R. Etienne. Em 1736, Shorter Oxford Dictionary registrou o uso da expressão inglesa "*thesaury or storehouse of knowledge*", definindo o termo como *"tesouro ou armazém de conhecimento, similar a um dicionário ou a uma enciclopédia"* (MOTTA, 2005). Através do relato dessa autora, fica clara a importância do trabalho de Roget, sempre citado na literatura de tesauros: o emprego do termo ganhou outro significado.

O final do século XIX e o início do século XX foram marcados pelo aparecimento das primeiras ferramentas de descrição de documentos com fins de pesquisa documentária, como a lista de cabeçalhos de assuntos utilizada no

[4] MIKHAILOV, A.I. Nociones generales acerca del tesauro. In: _____, Fundamentos de la informática. Moscú, La Habana, Nauka, Academia de Ciencias de Cuba. Inst. Document. E Inf. Cientifica y Técnica, 1973, v.2, p. 397-496.

catálogo da American Library Association em 1895 e a lista da Library of Congress em 1911. No domínio da classificação, destacam-se a Classificação Decimal de Dewey e a Classificação Decimal Universal. O primeiro trabalho pertence a Melvil Dewey e foi publicado primeira vez em 1876. Ele foi traduzido e ampliado por Paul Otlet e Henri Lafontaine, resultando na Classificação Decimal Universal cuja primeira edição foi em 1905 (CHAUMIER, 1978). O desenvolvimento da computação após 1960 possibilitou que os tesauros viessem a ser uma ferramenta efetiva em sistemas de documentação. É preciso lembrar, entretanto, que os trabalhos de computação se esquecem muitas vezes dos relacionamentos entre os termos e da própria teoria da informação, em função de uma otimização de processamento ou melhores índices estatísticos. Em Campos (2001) encontra-se uma retrospectiva histórica do tesauro na Europa e nos Estados Unidos.

Interessa-nos que o tesauro surgiu da necessidade de manipulação de grande quantidade de documentos especializados, onde é preciso trabalhar com vocabulário mais específico e uma estrutura mais depurada do que aquela presente nos cabeçalhos de assunto (remissivas e referências cruzadas tipo ‘ver’ e ‘ver também’). Parte do desejo de uma comunidade de usuários em recuperar documentos de uma área específica, onde é necessária maior sistematização para a recuperação. Ao descrever a intenção de sua obra, Bruschini et al. (1998) expõem claramente estes motivos.

O tesauro avançou na estrutura e nas referências cruzadas, dando lugar às relações hierárquicas (verticais) e associativas (horizontais). A grande característica de um tesauro é representada pelos termos e pelas relações entre eles. São estas relações que conferem a um tesauro uma multiplicidade

de usos, desde o auxílio na própria função de indexação até o auxílio na efetiva recuperação dos documentos.

Um vocabulário controlado constitui-se apenas de relações sinonímicas, ou quase sinonímicas, para controle da polissemia. Em um vocabulário controlado não se diferencia “termo” da “palavra” e não existem relações estruturais entre os elementos. Entende-se por palavra à “menor unidade léxica” e seu significado depende do contexto em que figura, pois tomada isoladamente pode ter diversos significados. Em um tesauro encontram-se termos, que possuem relacionamentos entre si, objetivando facilitar e propiciar melhor recuperação das informações que por ele foram indexadas. São relacionamentos dinâmicos e construídos pela própria evolução da área a que se refere o tesauro. É também uma forma de se organizar informações em determinado campo do saber e o aprendizado de seus termos dá-se naturalmente pelo próprio processo de utilização.

Elaborar um tesauro é antes de tudo uma atividade intelectual, que requer atividades específicas para a consecução dos objetivos dos que se empenham nesta tarefa, entre elas: o conhecimento de documentos produzidos na área, o entendimento dos termos empregados, a construção de conceitos para explicação dos termos. A construção de um tesauro requer uma atitude flexível para incorporar as mudanças que a linguagem utilizada sofre no caminho de seu desenvolvimento sem abrir mão dos conceitos, mas em atitude aberta a seu próprio desenvolvimento.

A ambigüidade que as palavras trazem, por carregarem em si mesmas múltiplos significados, faz com que se tornem inadequadas para a indexação e

a recuperação. As linguagens de documentação partem de uma palavra, tomada sob certos princípios, ou uma expressão, para representar um único conceito ou idéia. De antemão, pressupõe-se que estes conceitos ou idéias encontram-se legitimados na comunidade de usuários para a qual se constrói o tesauro.

Cada palavra, ou expressão recebe então o nome de TERMO, que equivale a um conceito adicionado de uma designação, construída por uma ou mais unidades léxicas. O controle é necessário para que se mantenha a relação de um termo para cada conceito e a cada conceito um só termo. Estes termos, que representam um conceito, são também chamados de descritores. Os outros recebem o nome de não-descritores e formam o conjunto das remissivas. Um tesauro é algo dinâmico, retrato fiel da realidade que representa, em contínua mudança, requerendo contínuas manutenções e atualizações. Seus componentes são: os termos, a estrutura entre eles e o conjunto de remissivas.

A estrutura de um tesauro diz respeito aos relacionamentos, às ligações e vinculações existentes entre os conceitos representados por termos. Nenhum termo existe em um tesauro sem ligação com outro, sempre determinada por seu significado (SVENONIUS, 2000). Os relacionamentos podem ser de diversos tipos, como:

- Relacionamentos lógicos, oriundos da comparação de dois conceitos. Podem se dividir em: genérico-específico, o que permite formar as classes dos conceitos e a estrutura hierárquica dos tesauros; relacionamento

analítico, que gera as relações a partir das relações associativas entre os termos; o relacionamento de oposição, indicando os termos opostos.

- Relacionamentos ontológicos, reunindo as relações partitivas, os relacionamentos de sucessão ou contigüidade entre os termos e os relacionamentos material-produto.

- Relacionamentos de efeito, constituído pelos relacionamentos de causalidade ou causa efeito, pelos relacionamentos de instrumentalidade e os relacionamentos de descendência que apresentam relações genealógicas entre os termos (GOMES, 1990).

Curras (1995) apresenta as relações existentes entre os termos de um tesauro classificadas em relações de equivalência, hierárquicas e associativas. Acredita-se ser esta uma divisão mais simples e mais próxima da hoje utilizada.

As relações de equivalência são aquelas que representam os sinônimos ou quase sinônimos de um termo. Na área técnica verifica-se a ocorrência de sinonímia com freqüência. Nos tesauros os sinônimos aparecem com a indicação de USE ou UP (use para). Um exemplo dá-se com as palavras: Antídoto e Contraveneno, que apareceriam nos tesauros um como termo preferido (ou descritor) e o outro fazendo remissiva ao primeiro.

Antídoto
UP Contraveneno
Contraveneno
USE ANTÍDOTO

A quase-sinonímia acontece quando dois conceitos têm praticamente a mesma intensão. Define-se intensão por um conjunto das características que

constituem um conceito. Na definição de um conceito, incluem-se em sua definição apenas as características mais importantes para o domínio do tesauro (BITI, 2005). Quando de sua ocorrência, seleciona-se um deles como descritor e procede-se como no exemplo acima, com uma entrada para cada um deles, estabelecendo uma relação de equivalência.

A homonímia é o emprego do mesmo termo com significados diferentes. A solução, quando ocorre, é a indicação do contexto. Exemplo:

Tênis (calçado)
Tênis (esporte)

As relações hierárquicas surgem da necessidade de especificar termos genéricos e específicos e desta forma, estabelecer os mais diversos tipos de relações, dependendo da linha e objetivo do tesauro.

Árvores
    Árvores frutíferas
        Macieiras
        Laranjeiras
    Árvores de nozes
        Nogueiras
        Castanheiras

Na bibliografia consultada assinalam-se as relações hierárquicas do tipo genérica, do tipo partitiva, ou seja, o todo com suas partes, do tipo enumerativa (Exemplo: Mar – Mar Báltico, Mar Negro, Mar Mediterrâneo) e as poli-hierárquicas, onde um termo pode depender de mais de um termo genérico.

As relações associativas são estabelecidas por pontos distintos de associação, dependendo do domínio do tesauro e de seu objetivo. O estabelecimento dos tipos de relações associativas faz parte da política de implantação do tesauro.

Resta observar que os tesauros são construídos para uma área específica do conhecimento. Não existe um tesauro geral; alguns cobrem vários assuntos. Os tesauros nascem da necessidade de se reunir e sistematizar a informação contida em documentos de determinado nicho do conhecimento. O trabalho de BRUSCHINI et al. (1998) é um bom exemplo. Assim, é possível definir um tesauro como uma "... *linguagem documentária dinâmica que contém termos relacionados semântica e logicamente, cobrindo de modo compreensivo um domínio do conhecimento*" (GOMES, 1990).

A função do tesauro é representar os assuntos dos documentos e das solicitações de busca. Esta representação é feita no momento da indexação, através dos processos consecutivos de análise do documento, identificação de seu conteúdo e da tradução para os termos do tesauro de acordo com a política de indexação. Na recuperação, a representação da solicitação é feita no momento em que o usuário busca uma informação, quando o pedido é analisado, identificando-se seu conteúdo. A seguir, busca-se o termo no tesauro através do processo de tradução. A própria estrutura do tesauro, ou seja, os relacionamentos nele existentes possibilitam este processo de tradução.

Seguindo este raciocínio, pode-se afirmar que o termo é um componente do sistema de recuperação de informação, sendo afetado por ele e diretamente afeta também seu desempenho. A estrutura de um tesauro é que traz ao

usuário – consulente ou indexador – a possibilidade de encontrar o termo mais adequado, mesmo se não se conhece o nome específico para a idéia ou conceito. A partir do que conhece o usuário, outros termos, oportunos ou não, são apresentados para sua escolha, através da estrutura do tesauro.

Gomes (1990) classifica primeiramente os tesauros em relação à língua suporte. Assim, eles podem ser monolíngues ou multilíngues. Uma atenção especial deve ser dada ao segundo tipo, pois na tradução de termos de um idioma para outro deve-se considerar o contexto cultural do emprego de determinados termos. Essa autora classifica também os tesauros pela especificidade de seus termos em macrotesauros e microtesauros. Nos macrotesauros os termos representam conceitos mais amplos, o número de descritores é pequeno e o número de remissivas é elevado, uma vez que conceitos específicos são representados por não-descritores, que remetem ao descritor genérico imediatamente superior. Nos microtesauros, os descritores estão atrelados a conceitos específicos e referem-se a uma área restrita do conhecimento.

Outra classificação feita pela autora diz respeito ao assunto: os tesauros podem ser voltados para uma missão ou problema, ou dedicados a um assunto. No primeiro caso, de tesauro multidisciplinar, um tesauro sobre meio ambiente é um bom exemplo; no segundo caso, de tesauros voltados para uma disciplina científica específica, um tesauro de Química constitui bom exemplo.

Alguns pontos específicos na teoria de Tesauros precisam ser ressaltados neste trabalho. São eles a garantia de usuário, também conhecida como garantia de uso, a garantia literária e a garantia estrutural.

2.4.1 Garantia literária, garantia de uso e garantia estrutural: bases para um tesauro

Uma das primeiras tarefas indicadas na bibliografia para a criação de um tesauro é a seleção de vocabulário, que vem logo após as atividades de planejamento do tesauro, ou seja, a definição do domínio e dos objetivos do instrumento, a seleção das fontes para consulta e a definição de sua forma de apresentação. Alguns autores indicam que o vocabulário deve ser levantado nas principais fontes do domínio do tesauro, indicadas na fase de seu planejamento, ou de algum vocabulário de um serviço de informação já consolidado da área (AITCHISON e GILCHRIST, 1979; BITI, 2005). O vocabulário selecionado para a criação de tesauros é, antes de tudo, um vocabulário normalizado, observa Svenonius (2000), pois é oriundo de um conjunto restrito de palavras e frases, que vieram da linguagem natural encontrada nas fontes do domínio e também, pelo tratamento semântico dos termos, para fixar o seu referente e estabelecer as suas relações com outros termos.

Nem todo vocabulário de um serviço de informação já existente deve ser aproveitado para um novo tesauro, pois nem sempre o conjunto como um todo interessa à clientela desse novo serviço (BITI, 2005). Selecionar o vocabulário é antes de tudo uma tarefa de delimitação do tesauro a ser construído, levando-se em conta principalmente as pessoas que dele farão uso e o próprio domínio do tesauro.

A definição do domínio do tesauro assegura as decisões da escolha do vocabulário a ser utilizado, bem como, as relativas aos termos a serem

adicionados ou removidos, o que colabora na questão do custo da construção de um tesauro, que está fortemente relacionado a seu tamanho. É importante limitar o tamanho de um vocabulário para se ter somente termos necessários e suficientes para alcançar seus objetivos (SVENONIUS, 2000).

O ideal seria que houvesse uma forma de conhecer precisamente se determinado termo deveria pertencer ou não ao vocabulário, o que é uma tarefa árdua e requer antes de tudo a experiência e o conhecimento da área de domínio do tesauro. Svenonius (2000) observa que, na prática, os domínios tendem a ser definidos indiretamente, especificando critérios para a seleção de termos, e que, tradicionalmente, os critérios empregados para este propósito são a garantia literária, a garantia do usuário e a garantia estrutural.

A garantia literária é um conceito introduzido em 1911 por Wyndam Hulme, autor que defendia que a determinação de classes utilizadas em linguagens de documentação não deveria originar-se da classificação do conhecimento, mas das classes existentes na literatura (DODEBEI, 2000). A garantia literária possui o *status* de um princípio: nas linguagens de indexação, o vocabulário escolhido para a representação dos assuntos deve ser derivado empiricamente da literatura para a qual pretende-se a representação. Ou seja, a literatura deve ser determinante.

Ao comentar as idéias de Hulme, Svenonius (2000) observa que a linguagem à qual ele se referia era a *Library of Congress Classification* (LCC), e a literatura que serviria como garantia era a contida nos livros que estavam na biblioteca. A autora acrescenta que a linguagem específica de uma área ou disciplina da ciência tem a propriedade de ser definida através de um texto canônico desta

disciplina ou de um conjunto de documentos da mesma. Uma vez que a literatura de uma área pode ser definida, as expressões e termos nela contidos são indicativos da temática e tornam-se candidatos à inclusão no vocabulário da linguagem.

O princípio da garantia literária é aplicado desde os processos de classificação. Para Vickery (1975) a classificação é antes de tudo uma ferramenta de seleção. Ao introduzir o processo de classificação, o autor enfatiza que o ato de agrupar termos em categorias deve estar acompanhado da observação de novos termos continuamente descobertos e cujas relações encontram-se gravadas na literatura. Para ele, não se pode ser demasiadamente rígido em apresentar antecipadamente um termo associado a uma categoria, pois esta tarefa também requer um exame concreto da literatura do assunto. A classificação deve ser baseada na garantia literária.

Embora necessária, a garantia literária não é suficiente para legitimar a admissão de termos no vocabulário de uma linguagem de documentação, pois nem sempre os termos utilizados pelos autores equivalem a termos utilizados por quem pesquisa e deseja recuperar a informação.

A garantia do usuário está relacionada com o princípio de que os termos selecionados para um tesauro precisam estar de acordo com aqueles utilizados pelos usuários na tarefa de recuperação de informação. Cutter (1904)[5] apud Svenonius (2000) já afirmava, naquela época, que o uso era o árbitro supremo, e ele não se referia aos escritores, mas àqueles que buscam e necessitam da

[5] CUTTER, Charles A., Rules for a Printed Dictionary Catalog, 4th ed. U.S. Bureau of Education Special Report on Public Libraries, Part II. Washington, D. C. U. S. Government Printing Office. 1904.

informação. Para Svenonius (2000), alguns teóricos da indexação por assuntos vêem a garantia do usuário competindo em importância com a garantia literária. Na realidade, ambas são importantes e complementares. É importante acatar o vocabulário dos usuários e através dele conduzir as requisições que eles fazem aos descritores de um vocabulário mais especializado. Mas, sabe-se também que muitas vezes alguns usuários se perdem em palavras no instante da busca. Para eles, é sempre útil incluir termos nos quais nunca poderiam pensar, mas para os quais poderiam ser dirigidos com a finalidade de melhorar os pedidos, em sua procura.

Comumente, a garantia do usuário é sempre honrada na criação de uma linguagem de documentação e é muito raro existirem tentativas contrárias a este princípio pelos especialistas nestas linguagens. Existem diversos trabalhos nesta área, inclusive trabalhos que pesquisam em registros de *logs* de transações feitas em bases de dados dispostas em tecnologia de bancos de dados *online*, onde é possível retomar a consulta de um usuário e tentar medir a adequação e o confronto do vocabulário por ele empregado nas rotinas de busca com os vocabulários normalizados de linguagens de documentação. Estes trabalhos demonstram que, no nível léxico, o encontro nem sempre se dá em nível significativo. Entretanto, no nível conceitual, em que os autores contabilizam os sinônimos e as relações genéricas, ele é consideravelmente alto. Tal fato fez com que Svenonius (2000) definisse que encontrar a linha de uso comum entre os termos nem sempre é o alvo principal para identificar os conceitos nos quais um usuário está interessado, mas é fundamental para dar nome a estes conceitos. Além de que, como reforça com propriedade a autora, são muitos usuários diferentes e com usos diferentes, e para traçar um

vocabulário normalizado é necessário incluir todos os nomes pelos quais um conceito é conhecido (SVENONIUS, 2000).

Também a garantia de uso já era observada muito antes. Gilchrist (1971) tratou a garantia de uso, de acordo com o contexto de sua época, como uma força paralela de pensamento na qual o leitor, como usuário e freqüentemente criador de literatura, deveria ser empregado como uma fonte para se conhecer a linguagem mais apropriada para a indexação. Ele cita Reisner[6] (1965) que já afirmava em seu trabalho ser essencial a presença do usuário, pois só ele, o usuário, seria capaz de refletir a diversidade semântica considerável da linguagem por ele utilizada. Gilchrist (1971) citando Greer[7] (1965), Papier e Cortelyou[8] (1962), continua seu raciocínio afirmando que os modelos que levam em conta o usuário estão diretamente relacionados com o potencial da necessidade de informação do usuário e com a linguagem por ele utilizada: uma linguagem "espontânea" com a qual expressa seu pedido de informação.

Lancaster e Warner (1993) partilham da mesma opinião em relação a ambas as garantias. Para eles, a escolha dos termos é decisiva para se construir um vocabulário controlado e o caminho principal para este processo deve ser necessariamente a escolha de termos derivados da literatura a ser indexada, incluindo no processo a garantia literária. Mas, há de se preservar a garantia do usuário, de quem deverão também vir os termos. São dois recursos que trazem um impacto na indexação e na busca e que provam particularmente a problemática da escolha de termos: (1) o nível apropriado de especificação do

[6] REISNER, Phyllis. Evolution of a "growing" thesaurus. New York. IBM Watson Research Center. 1965.
[7] GREER, F. User vocabulary in thesaurus development. Perceptual and Motor Skills, v. 21. p. 827-37. 1965.
[8] PAPIER, L. S. , CORTELYOU, E. H. Use of a technical word association test in the preparation of a thesaurus. Journal of Documentation. v.18, n.4, p. 183-7. 1962.

vocabulário e (2) a extensão às palavras, que são combinadas para dar forma a termos mais complexos e mais específicos da indexação.

Acontece muitas vezes, que alguns termos não se apóiam nem na garantia literária nem na garantia de uso, mas são admitidos em um vocabulário normalizado porque possibilitam uma função estrutural útil. São termos que facilitam elos em uma hierarquia de termos ou colaboram para que seja possível dispor um conjunto mais específico de termos.

Quando se constrói um vocabulário hierárquico, o mais provável é admitir termos que estão somente garantidos por suas propriedades estruturais. Vocabulários altamente estruturados, como os empregados por linguagens classificatórias, incluem muitos assuntos designados por expressões que dificilmente alguém pensaria em usar por escrito em documentos ou em arquivos de tecnologia, isto é, bancos de dados de aplicações automatizadas. Svenonius (2000) exemplifica o fato citando títulos encontrados na Classificação Decimal de Dewey que começam com a palavra “tipos”, como “Tipos de escolas”. Embora nem sempre requeridos, estes títulos existem em virtude da sua utilidade nas funções de navegação pela estrutura.

A isto denomina-se garantia estrutural, ou seja, termos cuja colocação encontra justificativa na estrutura do tesauro. Geralmente, para esses termos, a garantia estrutural tende a anular a garantia literária ou a garantia de uso e a sua colocação resulta em uma seleção de vocabulários da forma *top-down* ao invés de uma seleção *bottom-up* (SVENONIUS, 2000).

As relações existentes entre os termos de um tesauro constituem ponto fundamental de tal forma que as diretrizes de alguns tesauros proíbem termos

órfãos, isto é, não relacionados a outros termos. O que faz sentido, se o propósito da normalização de vocabulário é estabelecer conectividade na forma de relações semânticas entre termos. Também é significativo se o vocabulário for usado junto com texto livre para pesquisa. Seria uma despesa desnecessária incluir termos órfãos que poderiam ser capturados através de palavras-chave. Por outro lado, onde o vocabulário de assunto é o único acesso à informação, baseado em índice impresso, é necessário incluir termos órfãos ou conceitos importantes para que possam ser recuperados.

2.4.2 A garantia que vem das unidades de texto

Algo particular chamou a atenção ao se construir este projeto. Os textos científicos obedecem a determinado formato, com normas para sua construção. E algumas partes específicas de um texto são prioritárias tanto para quem escreve quanto para quem os lê. Passemos a definir um texto científico publicado no formato de um artigo.

Um artigo científico possui uma estrutura composta de elementos pré-textuais, textuais e pós-textuais. Os elementos pré-textuais compõem-se do título, da autoria do artigo, do resumo e das palavras-chaves. Os elementos textuais correspondem ao texto propriamente dito e os pós-textuais às referências bibliográficas, aos resumos em outras línguas, às notas de final de texto e anexos. Existem diversas obras que tratam da normalização de trabalhos científicos. Entre elas foram vistas as obras de França et al.(2004) e Cunha (2004). No caso brasileiro, as normas técnicas da Associação Brasileira de Normas Técnicas (ABNT) regulamentam a apresentação desta estrutura e das

particularidades de cada um destes elementos através das normas técnicas NBR10520, NBR6023, NBR6022, NBR6024, NBR6028.

As normas da ABNT dão diretrizes para uma correta apresentação dos elementos de um artigo científico. O título deve ser claro e objetivo, podendo ser formado de título e subtítulo, apresentado sempre na mesma língua do texto e, em caso de necessidade de apresentação em outras línguas, estes são apresentados logo a seguir. No título de artigo científico, deve-se evitar a utilização de parênteses e fórmulas que dificultam a compreensão de seu conteúdo. Se o artigo for uma tradução, o nome do tradutor e o título original do artigo devem ser apresentados em notas de rodapé.

Segundo Krzyzanowski e Ferreira (1998) foi a partir da década de 60 que começaram a surgir na literatura estudos sobre avaliação de revistas científicas e técnicas, demonstrando a necessidade de se definirem parâmetros mensuráveis que realmente refletissem a qualidade da informação registrada. Os autores fazem referência a um artigo publicado por Arends[9] (1968), onde é relatada uma avaliação dos periódicos médicos venezuelanos, baseada em modelo criado por um grupo de trabalho da Unesco em 1964, para a seleção de revistas técnicas latino-americanas. Em 1982, Braga e Oberhofer[10] apud Krzyzanowski e Ferreira (1998) apresentaram uma proposta para avaliação de periódicos brasileiros científicos e técnicos, alterando o modelo da Unesco. Esses autores propõem um modelo que procura refletir aspectos de forma dos periódicos, dentro de parâmetros mensuráveis.

---

[9] ARENDS, L. Las revistas médicas venezolanas: evaluación de su cualidad. Acta Cient. Venezolana. V. 19, p. 145-151, 1968.

[10] BRAGA, G. M., OBERHOPER, A. Diretrizes para avaliação de periódicos científicos e técnicos brasileiros. Ver. Lat., n. 1, p. 27-31. ene./jun., 1982.

Lopes Neto et al. (2002) apresentam pesquisa realizada em títulos de artigos de um periódico de enfermagem. Este trabalho merece atenção, já que o problema apresentado por eles focava-se na adequação dos títulos das pesquisas de enfermagem em traduzir o que realmente havia sido estudado. Abaixo são listadas algumas das conclusões e observações feitas por estes autores, que podem ser estendidas a artigos de periódicos de outras áreas da ciência:

- Os títulos, de forma geral, possuíam linguagem econômica em palavras como uma condicionante de objetividade e veracidade;
- Os títulos deviam oferecer aos leitores de pesquisas científicas contribuições de forma direta, facilitando sua busca sem gerar ansiedade e frustração;
- Títulos exuberantes, construídos na tentativa de mostrar uma linguagem filosófica e científica, comumente deixam de expressar o conteúdo trabalhado. O uso de uma linguagem rebuscada pode decepcionar os leitores, quando o ideal seria motivá-los;
- A maioria dos artigos analisados apresentava títulos considerados adequados, de acordo com o referencial de análise utilizado, embora uma quantidade significativa de artigos apresentasse títulos com uma captação apenas parcialmente adequada, uma vez que, de acordo com a percepção dos autores, não condiziam com o conteúdo expresso no corpo do trabalho;
- O emprego de metáforas em títulos foi verificado em apenas 7,35 % dos casos e o pleonasmo em 5,38% dos mesmos;

- Títulos longos foram encontrados em 16,37 % dos artigos e títulos com problemas de pontuação - emprego inadequado de sinais de pontuação como interrogação, dois pontos, etc. - foram observados em 13,24 % dos títulos analisados;

- Em relação à clareza do título, esta foi observada em 66,17% dos casos;

- As palavras-chave utilizadas nos artigos foram parcialmente captadas nos títulos e, segundo os autores, deve-se dar uma atenção maior a este emprego e estudo.

Nos textos científicos, obrigatoriamente o nome do autor, ou dos autores, é sempre apresentado por extenso e a seguir ao título, embora em alguns casos seja possível ver nomes intermediários abreviados. Se o periódico exigir a apresentação das credenciais do autor, estas devem ser apresentadas em nota de rodapé. Agradecimentos dos autores e data de entrega dos originais à redação, devem aparecer em nota editorial no final do artigo.

O resumo é uma apresentação concisa e seletiva de um texto, apresentado sempre em seqüências de frases afirmativas e não de enumeração de tópicos, ressaltando o objetivo, o método e, de forma mais genérica, os resultados e as conclusões do documento. O resumo é sempre disposto na língua do texto e preferencialmente, deve-se utilizar o verbo na voz ativa e na terceira pessoa do singular. No resumo *"... a primeira frase deve ser significativa, explicando o tema principal do documento..."* (NBR6028). Quanto à sua extensão, os resumos de artigos de periódicos não devem ultrapassar a 200 palavras, exigência não aplicada quando se trata de resumos críticos de obras.

Seguindo ao resumo, são apresentadas as palavras-chave na língua do texto, consideradas um elemento obrigatório do texto científico, salvo alguma recomendação específica do periódico, em contrário são antecedidas pela expressão “Palavras-chave:”, separadas entre si por ponto e finalizadas também por ponto (NBR6022) Correspondem a palavras significativas do conteúdo do artigo e facilitam a elaboração posterior de um índice de assunto. França et al. (2004) recomendam que estas palavras devem ter uniformidade de apresentação em todo o periódico, consistência de estilo e exatidão de informação em relação ao texto científico. Sempre que possível, devem apresentar exaustividade, ou seja, devem ser o mais completo possível, respeitando o número de palavras-chave limite do periódico. E é fundamental que sejam claras e adequadas ao leitor a quem se destinam.

Logo depois é apresentado o texto, a parte principal do artigo, contendo uma exposição ordenada e pormenorizada do assunto tratado. Divide-se em seções e subseções, cuja regulamentação encontra-se nas normas técnicas NBR6024 e NBR6022. Via de regra, encontram-se nele alguns elementos de destaque: a introdução, o desenvolvimento do trabalho contendo uma revisão de literatura através dos trabalhos relacionados, referencial teórico, a metodologia empregada, os resultados e a avaliação destes e a conclusão. Estes itens conterão sempre sub títulos e numeração conforme o padrão do periódico onde se publica o texto científico. Atenção especial é necessária para as citações, regulamentadas pela norma NBR10520 da ABNT.

As referências bibliográficas constituem um elemento obrigatório em textos científicos e devem ser apresentadas conforme a norma NBR6023. Correspondem a uma listagem de todas as obras e documentos bibliográficos

que foram usados para elaboração do artigo. Em alguns casos, após as referências bibliográficas são apresentados as notas de final de texto, os agradecimentos e os anexos.

Wilson (1985) trata do assunto do texto. Para ele um autor, ao escrever, tem uma intenção clara quanto ao seu assunto. De alguma forma, passa a ter controle sobre nosso sentido em relação ao domínio do texto, que ele, autor, controla muito bem. Ao escrever, sabe o que está fazendo e fornece um enunciado honesto em relação a seus objetivos através do próprio texto, o que favorece àqueles que trabalham na representação do texto.

Neste trabalho, o título de um artigo científico, o resumo, as palavras-chave e a bibliografia tornaram-se objetos de estudo, em particular pelo entendimento de que a estas partes de um texto científico são dadas prioridades de construção, pois são instrumentos que facilitam a escolha do usuário na recuperação. Elas passam a ser as unidades de texto eleitas para a pesquisa. Assim sendo, levanta-se a questão de que a partir delas é possível extrair termos e relacionamentos representativos de um documento, de uma coleção de documentos e de um conjunto de coleções de documentos.

2.5 Alguns trabalhos relacionados

Lancaster e Warner (1993) apresentam comparações entre trabalhos realizados na Ciência da Informação sobre o tema de construção de tesauros. A pesquisa nesta área encontra-se dividida em duas grandes linhas: uma de construção manual de tesauros e a outra de construção automática de tesauros. Nessas pesquisas, o que está sendo representado é a organização

de uma terminologia específica de assuntos e o objetivo de ambas as linhas é impor maior predição no uso da língua, na recuperação de informação, pela representação explícita de relacionamentos entre termos e frases. Este procedimento colabora para que a revocação seja acompanhada de alto nível de precisão, tendo em vista a especificidade da terminologia representada – "*quanto mais específicos os termos encontrados no tesauro, maior a precisão*" (LANCASTER e WARNER, 1993).

Na construção manual, palavras e frases são identificadas por um grupo de peritos em documentação, através da extração de assuntos da literatura. Estes assuntos são então organizados em facetas e são determinadas as relações entre os termos. As pesquisas na linha de construção automática dividem-se em dois grupos: trabalhos que partem de estatística de ocorrências de termos e trabalhos que têm como base a lingüística.

Soergel (1974) aponta que os relacionamentos extraídos através da observação estatística são relacionamentos conceituais e esta é a idéia básica para métodos automáticos de construção de tesauros. Para ele, os métodos automáticos auxiliam o processo de construção, mas não substituem o esforço intelectual necessário para a construção de uma linguagem de indexação ou mesmo de um tesauro. As estatísticas não devem ter precedência ao julgamento humano na avaliação do vocabulário, mas através delas algumas decisões úteis podem ser tomadas. O autor considera que a identificação dos termos e dos relacionamentos por métodos automáticos deve ser considerada como um tipo e pré-processamento das fontes, especialmente sumários e textos completos, antes de se decidir transferir termos para o tesauro (SOERGEL, 1974).

Na aproximação automática existem sempre dois procedimentos: a seleção prévia através da utilização de freqüência, eliminando anteriormente *stop words* (palavras sem significado para a indexação ou representação do documento); seguindo com a seleção de termos que representam estes conteúdos. O trabalho de Luhn (1997) e de Salton (1981) são representantes desta linha. Hirschman, Grishman, Sager[11] (1975) apud Lancaster e Warner (1993) utilizam-se também da classificação dos termos em classes gramaticais para a seleção. Lancaster e Warner (1993) citam também os trabalhos de Dilton e McDonald[12] (1983) onde frases mais específicas e complexas são utilizadas para análise lingüística, assegurando uma seqüência gramatical.

Alguns outros trabalhos agrupam palavras ou frases estatisticamente, de acordo com a freqüência de suas co-ocorrências ou através de sua freqüência relativa, como o de Spark Jones (1971)[13], Stilles (1961)[14] e Doyle (1961)[15] todos eles citados por Lancaster e Warner (1993).

Soergel (1974) coloca que trabalhos estatísticos podem ser considerados em unidades de texto, equivalendo às partes de formulações de pedido de buscas tendo em vista termos dos descritores; ao conjunto dos termos indexados contidos na representação do documento; aos resumos de documentos; às frases individuais extraídas dos documentos ou de seus resumos; aos parágrafos dos resumos ou mesmo ao texto completo. Esse autor denomina

---

[11] HIRSCHMAN, L., GRISHMAN, R., SAGER, N. (1975). Grammatically-Based Automatic Word Class Formation. *Information Processing and Management v.*11, p.39-57.
[12] DILTON, M., MCDONALD, L. Fulty Automatic Book Indexing. Journal of Documentation, v. 39, n. 3; p.153-154, september 1983.
[13] SPARK JONES, K. Automatic keyword classification for information retrieval. London, Butterworth, 1971.
[14] STILES, H. E. The association factor in information retrieval. Journal of the Association For Computing Machinery. V. 8, n.2, p.271-279, April 1961.
[15] DOYLE, L. B. Semantic road maps for literature searches. Journal of Associations for Computing Machinery, v. 8, n. 4, p. 553-578, Octobre 1961.

"*corpus do texto*" ao conjunto das unidades de um texto de um ou todos os tipos. Para ele existem dois métodos principais de se contar freqüência de termos: uma contagem denominada total, equivalente ao número de vezes que o termo ocorre no documento ou uma contagem única para cada unidade, independente do número de ocorrências do termo. Aponta, também, a possibilidade de se estabelecer pesos de acordo com a freqüência do termo dentro da unidade.

Soergel (1974) discute ainda a validade de serem os descritores retirados de um texto apenas por alto valor de ocorrência dos mesmos observando que quando aparecem em poucas unidades eles podem ter maior poder discriminatório. Em seu trabalho, o autor apresenta formas de mensurar a ocorrência simultânea ou não de dois termos em unidades do texto, assim como uma interpretação destes valores de ocorrência, tendo em vista a possibilidade de existir relação de sinonímia, de homonímia, relação conceitual ou uma relação hierárquica entre estes termos.

Conforme Lancaster e Warner (1993), o que não aparece nas aproximações estatísticas é uma representação das relações semânticas específicas entre os termos, o que pode levar a agrupamentos "*ambíguos semanticamente"* que, gerados deste modo, deixam questionamentos sobre a utilização de termos extraídos a partir de uma coleção em outra coleção distinta.

Para operacionalizar este problema outros autores tentaram produzir um tesauro como uma rede semântica, através do uso de um dicionário legível por

computador. Fox et al.[16] (1998) apud Lancaster e Warner (1993) apresentaram trabalho nesta linha. Um vocabulário controlado gerado deste modo presumivelmente não estaria unido a uma coleção particular e representaria o conteúdo de um vocabulário geral, passível então de aplicação em muitos ambientes diferentes.

Existem também tesauros construídos por recursos léxicos genéricos, como os dicionários que são potencialmente úteis, já que o conjunto de termos e frases estaria relacionado de forma semanticamente específica, como por exemplo, através de relações gênero-espécie, parte-todo, causa-efeito, etc. Esse tipo de processamento requer um alto desempenho computacional (LANCASTER e WARNER, 1993), porém construir tal rede é necessário, já que ainda não está claro como um tesauro assim construído colaboraria nas operações de recuperação. Esses dicionários, em geral, partem da estrutura da língua, sem a necessária presença da terminologia especializada, requerida nas solicitações de busca de assuntos específicos.

Lancaster observa ainda que o uso da Lingüística, como base para se agrupar termos, pode significar melhoria de desempenho em alguns casos de recuperação de informação, conforme apresentado por Wang, Vandendorpe e Evens (1985)[17] *apud* Lancaster e Warner (1993). Estes autores propõem a ruptura de termos específicos e termos genéricos em algumas categorias de relacionamento, como nas relações parte-todo, como forma de melhoria em alguns casos de recuperação de informação.

[16] FOX, E. A. et al. Building a large thesaurus for information retrieval. In: Proceedings of the Second Conference on Applied Natural Language Processing. Austin, Texas, February 1998. Morristown, N. J., Association for Computational Linguists, 1988, p. 101-108.

[17] WANG, Yih-Chen, VANDERDORPE, J., EVENS, M. Relational thesauri in information retrieval. Journal of the American Society for Information Science. v. 36, n.1, p.15-17, January 1985.

Na Ciência da Computação os trabalhos encontrados versam sobre a automação da construção de tesauros, sendo que alguns trabalhos relacionados com processamento de linguagem natural articulam a possibilidade de se trabalhar com palavras-chave, vocabulário controlado e tesauros. No contexto deste último, a computação vem sendo utilizada para a construção do vocabulário a ser utilizado em um tesauro, em comparações de semelhanças entre palavras ou termos empregados em uma área, através de associações e estatísticas, e para a própria organização do vocabulário.

Não se pode deixar de referenciar o desafio para recuperação de dados na Web e as contribuições da computação através das máquinas de busca. Shiri e Revie (2000) apresentam os tipos de tesauros existentes para a Web, delineando suas funções e os desafios a serem vencidos. O trabalho de Chen et al. (2003) propõe a construção automática de tesauros para a Web, através da estrutura de vínculos (*links*) encontrada nas páginas. Está presente também a preocupação com as atualizações de termos de determinado domínio, visto que a proposta é identificar novos termos e reconhecer novas relações para um tesauro, através da própria evolução da rede. Os autores partem de um conjunto de páginas da Web representativo do domínio do tesauro. A metodologia consiste em filtrar os vínculos dispostos, analisando cada conteúdo das páginas apontadas, as quais são representadas por um grafo onde cada nodo representa um conceito. As relações trabalhadas são de associação e agregação. As relações associativas são representadas por relações horizontais neste grafo, sendo que as relações de agregação compõem a estrutura hierárquica. O tesauro é então construído pelas

estruturas de conteúdos das páginas selecionadas. Estes autores chegaram a uma melhoria de 20% na busca por conteúdos contidos nestes tesauros.

O trabalho de Crouch e Yang (1992) apresenta um algoritmo para geração automática de tesauros através da construção de vetores representativos de estatísticas das relações existentes entre as palavras. Os autores validam o algoritmo apresentando resultados de melhorias na efetividade de recuperação em quatro coleções utilizadas como massa de teste. Nestes dados, a pertinência de um termo a uma classe do tesauro é identificada através de algoritmos. Kimoto e Iwadera (1989) mostram um sistema de recuperação baseado em um tesauro dinâmico utilizando os conceitos de conexão entre nodos, com base em uma rede de estrutura onde cada nodo representa um termo do tesauro. As conexões entre esses nodos representam as ligações, ou relações, entre os termos e, por sua vez, entre os documentos. Os autores previram a informação de documentos relevantes para o usuário, abrindo possibilidades de alterar o peso do nodo e gerar vínculos. De alguma forma, os pesos de nodo e vínculos refletem um interesse particular do usuário.

Bruandet (1982) apresenta uma metodologia em forma de estudo teórico para construção automatizada de tesauros através da utilização de bancos de dados, e Donghong (1998), um estudo para combinar um tesauro com um dicionário, ambos na língua chinesa. As palavras do tesauro são localizadas no dicionário de onde são capturadas as palavras relacionadas. Ou seja, trabalha com relações de equivalência em tesauros, a partir de seus termos em comparação com as palavras de um dicionário. O trabalho produz uma classificação semântica das significações encontradas no dicionário. Foo et al. (2000) trabalham também na construção de um tesauro automático chinês;

explicitam as fases de construção de um tesauro automático, enquanto apresentam suas experiências trabalhando com um conjunto de documentos da área de economia.

Braga (1982) apresenta estudo realizado para indexação de artigos de periódicos, a partir de palavras extraídas dos títulos, utilizando algoritmo *keyword-in-context* - KWIC - aplicado a cada título e mensurando estatisticamente os resultados. Para a autora, o título é uma forma de sumarização que determina o conteúdo do documento, ou de outra forma, sobre o que o documento foi escrito e por isso justifica esforços de indexação a partir do mesmo. O KWIC foi introduzido por Peter Luhn em 1958 e baseia-se na rotação automática das palavras que compõem o título. Na verdade, o método encontra-se apoiado em três princípios básicos:

- os títulos são informativos;
- as palavras extraídas dos títulos podem ser usadas para uma indexação e uma busca efetiva;
- a palavra vista isoladamente pode ter sentido ambíguo e o contexto que a rodeia permite definir e explicar o seu significado.

A autora registra a cifra de 70 artigos encontrados desde 1958 e 1960 (datas das obras de estudo de Luhn) até 1980, que tratam de título como elemento de aceso ao conteúdo. Kremmer (1985) aponta que os índices gerados pelos algoritmos KWIC, ou *keyword-out-context* (KWOC), são limitados em recursos de busca. Esses índices são ainda mais limitados em recursos de busca, mas são importantes quando se deseja fornecimento rápido de informação sobre a

literatura corrente de determinado campo. Joyce e Needham (1997) apontam que Luhn acreditava que o método estatístico poderia ser aplicado da seguinte forma:

- palavras com significado semelhante ou relacionado se agrupam em famílias, como nos cabeçalhos em Roget's Thesaurus;

- a codificação de documentos em termos de elementos nocionais é efetuada por meio do dicionário de anotações e o resultado é um abstrato nocional mecanicamente preparado;

- para recuperação, é solicitado ao usuário da pesquisa preparar uma composição com tantos detalhes quantos os do problema. Isto é codificado da mesma maneira, e o padrão de pergunta nocional é comparado com o padrão notacional do documento. Uma vez que um *match* idêntico é altamente improvável, este processo poderia levar a uma base estatística partindo de determinado grau de semelhança.

Chen e Wu (1999) afirmam que a qualidade de indexação não depende somente de conhecimento profissional e da experiência de bibliotecários ou especialistas de assunto; pode ser afetada por restrições impostas em relação ao tempo e ao custo da atividade. Em seu trabalho, partem de palavras de títulos e resumos de documentos indexados manualmente como treinamento para a criação de um vocabulário controlado a ser gerado. A idéia proposta por eles é baseada no conteúdo das palavras do vocabulário, presumindo-se que deve existir uma relação entre vocabulários controlados e o conteúdo das palavras extraídas. Ao se encontrar estas palavras nos textos, eles atribuem ao

documento a indexação do vocabulário controlado, permitindo que um mesmo texto seja indexado por diversos assuntos.

Morita et al. (2004) trabalham com a construção automática de tesauros. Em pesquisa relativa à co-ocorrência de palavras, através do processamento de linguagem natural, utilizaram-se de um dicionário na língua dos documentos e, a partir de hierarquias construídas com as palavras dos documentos e com as do dicionário, extraíram estas co-ocorrências. Nesse trabalho, apresentam cálculos para se verificar a similaridade entre as palavras. Hodge e Austin (2002) apresentam a hierarquia léxica de palavras através de conceitos ligados a redes neurais. Seu trabalho descreve uma metodologia para categorização de palavras dentro destes princípios.

Holanda et al. (2004) utilizam-se de grafos construídos entre as palavras de um tesauro, com interesse pela forma como se dá o caminho entre as palavras, ou seja, entre os nodos do grafo. Neste, as conexões entre os nodos correspondem às relações entre os termos do tesauro.

Salton (1981) apresenta um sumário das possibilidades de indexação através de aplicações automatizadas por freqüência de termos. Para ele, existe um relacionamento claro entre a freqüência de um termo e sua importância para representação do conteúdo do documento. Aponta estas questões tendo em vista o documento e uma coleção de documentos, levantando pontos sobre a especificidade de um termo e seu poder de discriminação de documentos.

3 O ambiente desenvolvido

Este capítulo apresenta o caminho da construção do ambiente semi-automatizado para a construção de tesauros como alternativa de solução para o problema levantado. Apresenta-se a seguir a fase de pré-teste e os passos para a construção do ambiente desenvolvido.

3.1 Pré-teste: verificando possibilidades através dos dados

Como pré-teste, foi realizado um experimento para verificar se a construção de uma base de dados, formada pela reunião de todas as palavras-chave, encontradas em textos específicos de uma área, aumentaria o poder de indexação dos mesmos para uma máquina de busca baseada em vocabulário controlado. Partiu-se da hipótese de que as palavras-chave informadas por autores nos textos científicos ocorrem nas partes específicas: título, resumo, texto e bibliografia, neste trabalho tratados como unidades de texto.

O experimento foi realizado em duas fases: na primeira, foi verificado se as palavras-chave de cada artigo eram encontradas em cada uma dessas unidades; na segunda fase, foi criado um arquivo contendo todas as palavras-chave de todos os artigos e novamente foi realizada a verificação de ocorrência de palavras-chave nessas unidades, partindo-se agora do conjunto de palavras-chave.

Como massa de teste foram escolhidos 114 artigos escritos na língua portuguesa e disponibilizados no periódico digital *Datagrama Zero*[18] em julho

[18] Na bibliografia a referência DATAGRAMA ZERO(2004) consta com data de seu ultimo acesso.

de 2004, quando foi realizado o pré-teste. Nessa fase, os resultados foram obtidos a partir de uma análise quantitativa dos dados levantados, com execução de programas para reconhecer a ocorrência de palavras-chave nos textos do periódico.

As Figuras 1 e 2 apresentam o percentual das ocorrências de palavras-chave nos títulos. Apenas 27% dos artigos da massa não possuíam suas palavras-chave em seus títulos. Este percentual passa para 10% dos artigos quando estas palavras são comparadas com o vocabulário controlado criado a partir do conjunto de palavras-chave dos artigos. Pode-se supor que o uso do vocabulário formado pelo conjunto das palavras-chave encontradas nos artigos para indexar seus títulos levaria a um ganho nas possibilidades de recuperação da informação.

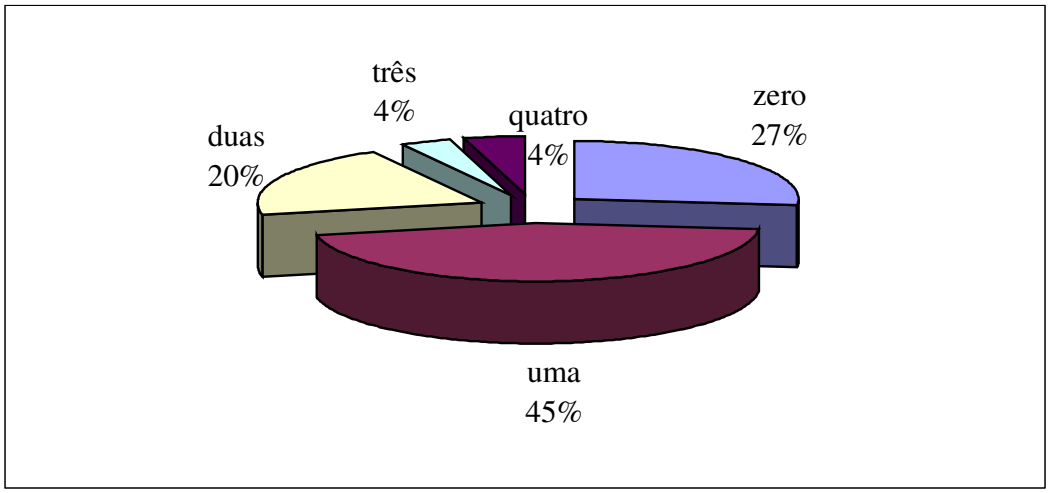

Figura 1 - Percentual de artigos com a quantidade de palavras-chave encontradas em seus títulos

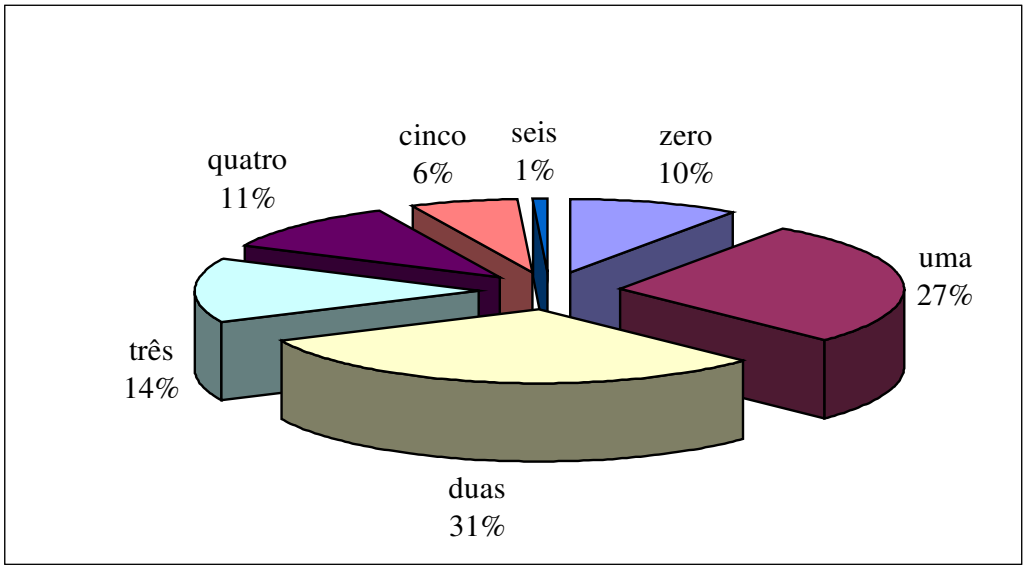

Figura 2 - Percentual de ocorrência de palavras-chave da base nos títulos de artigos

As Figuras 3 e 4 apresentam o percentual das ocorrências de palavras-chave nos resumos. Dos 114 artigos da amostra, 24 (21%) não incluíam suas palavras-chave nos resumos. O vocabulário controlado reduziu este número para apenas dois artigos, pouco menos que 2% do total. O vocabulário controlado aplicado aos resumos pode ser eficiente na recuperação de informações; mas prevalece a premissa do conhecimento deste vocabulário e sua aplicação pelo usuário.

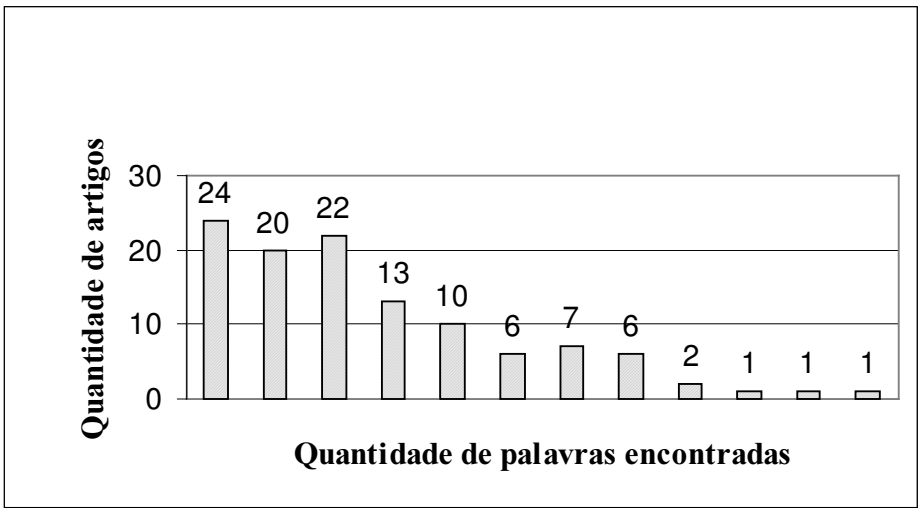

Figura 3 - Artigos com suas palavras-chave no resumo

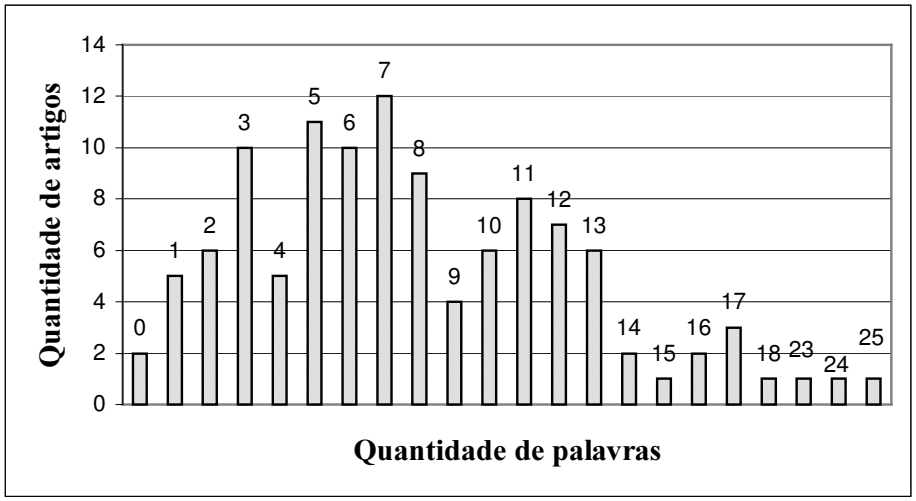

Figura 4 - Artigos com palavras-chave da base no resumo

O resultado dos testes para encontrar palavras-chave nos textos dos artigos ficou dentro do esperado. Apenas oito artigos, 7% do total, não possuíam as palavras-chave no corpo do texto. Um dado interessante é que seis deles não

possuíam também estas palavras no título e na bibliografia. A aplicação utilizando a base de palavras-chave reduz este valor a zero.

Existem hoje na Web processos que procuram demonstrar quantas vezes um texto já foi citado em outros textos. Um exemplo é o *Citeseer.Ist – Scientific Literature Digital Library*[19]. Nesta fase de testes partiu-se da hipótese de que a bibliografia utilizada pelos autores dos artigos poderia apresentar termos coincidentes com as palavras-chave do artigo, auxiliando a recuperação da informação. Apesar de quarenta artigos (35%) não apresentarem palavras-chave de seus textos em sua bibliografia, o percentual dos que apresentam é animador.

O emprego da base de palavras-chave em relação à bibliografia reafirma estes dados. Dos quarenta artigos iniciais, cujas palavras-chave não encontraram correspondentes na bibliografia, apenas sete continuam desta forma ao se aplicar o vocabulário controlado. Este número é muito pequeno e representa apenas 6% do total dos artigos, evidenciando que a forma de recuperação por bibliografia pode também facilitar a recuperação de informação. As Figuras 5 e 6 ilustram esta execução.

[19] CITESSER.IST. Scientific Literature Digital Library. Disponível em http://citeseer.ist.psu.edu Acesso em 10 jul. 2004.

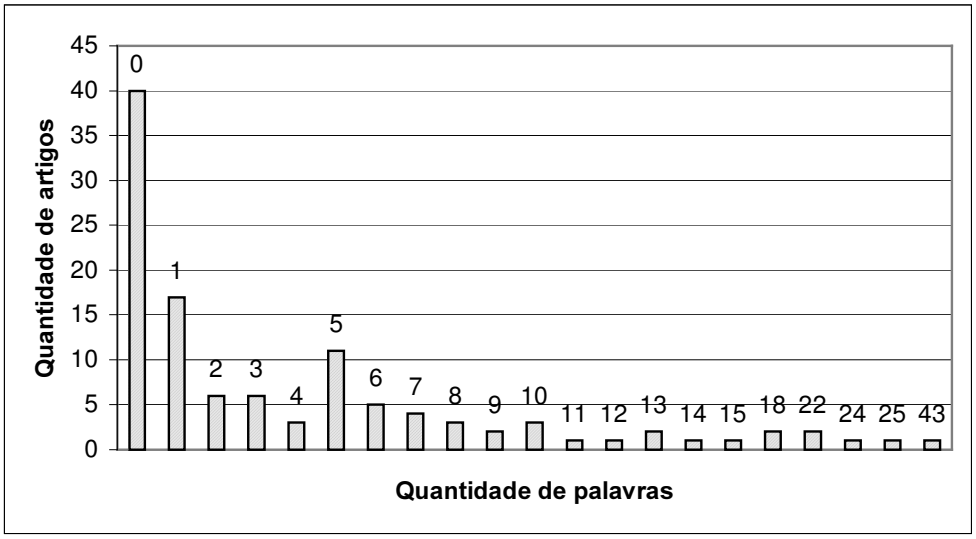

Figura 5 - Quantidade de palavras-chave encontradas nas bibliografias dos artigos

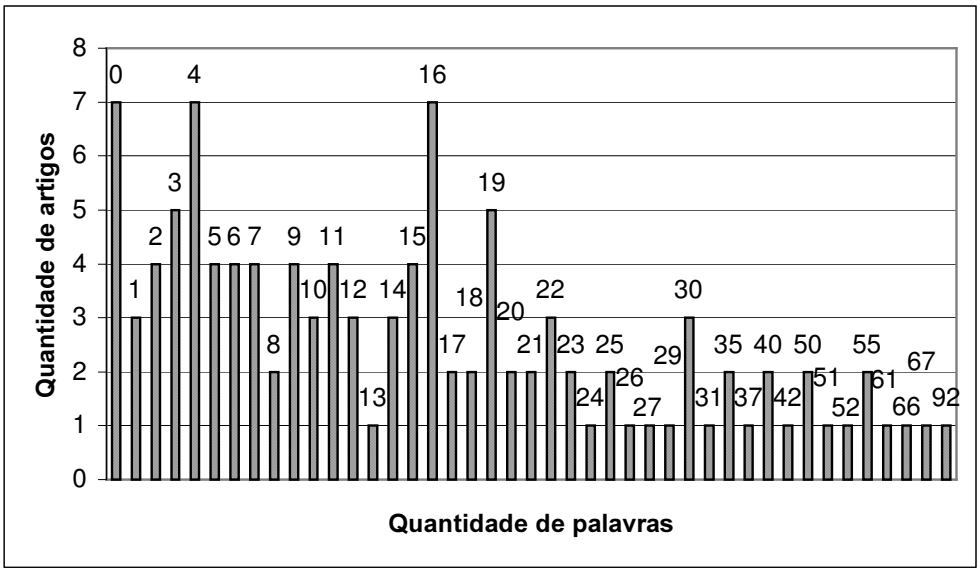

Figura 6 - Quantidade de palavras do vocabulário encontradas na bibliografia

Assim, nesta fase de pré-teste foram verificadas as ocorrências das palavras-chave indicadas pelos autores de textos científicos de determinada área do conhecimento no título, no resumo, no próprio texto e na bibliografia dos artigos da revista Datagrama Zero. Os resultados do pré-teste demonstraram que, na maioria dos casos, essas palavras facilitam a recuperação dos artigos, se os

mesmos forem indexados pelo conjunto das palavras. Os resultados do pré-teste incentivaram a continuidade da pesquisa e tornaram-se base para a formulação da hipótese deste trabalho.

3.2 AlfaThes: ambiente semi-automatizado de geração e alteração de tesauros

Entende-se por ambiente de tecnologia a reunião de recursos de máquinas (*hardware*), programas (*software*) e, principalmente, de pessoas que executam tarefas e interagem com o mesmo. O ambiente semi-automatizado de construção e alteração de tesauros foi batizado - e será assim denominado neste trabalho - com o nome de *Alfathes* (Alfa Thesaurus) pela crença de que, com ele, apenas inicia-se um caminho, em analogia à letra grega alfa, início de tudo.

Para a programação do AlfaThes utilizou-se a linguagem PHP (acrônimo recursivo para "PHP: *Hypertext Preprocessor*"), em sua versão 4. O PHP é uma linguagem voltada para o desenvolvimento de aplicações Web. A escolha deu-se principalmente pelas funções que possui para gerenciamento de *string* (conjunto de letras), facilitando o processamento a ser realizado. Para armazenamento dos dados optou-se pelo gerenciador de banco de dados MySql, software livre, que utiliza a linguagem SQL (*Structured Query Language*) como linguagem de manipulação de dados (DML – *data manipulation language*). Uma das principais características do MySql é sua total integração com o PHP, sendo que a interface utilizada para consultas no projeto foi o EasyPHP (EASYPHP, 2005).

Foram selecionados como usuários deste ambiente pessoas com um conhecimento profundo em criação de linguagens de indexação, aqui denominados especialistas em indexação, e usuários com conhecimentos específicos na área de domínio do tesauro, aqui denominados especialistas da área. Ambos os tipos de usuário serão referenciados neste texto como usuários gestores do ambiente, o qual contou também com um gestor, aqui denominado gestor do projeto, cujo perfil é semelhante ao dos demais usuários.

Foram tomadas como premissas básicas que o ambiente deveria possuir interface amigável e possibilitar a todos os usuários envolvidos a consulta a todo e qualquer dado armazenado em sua base: a coleção de documentos, as palavras-chave extraídas, as palavras-chave tratadas, os termos recomendados, os conceitos construídos, os termos e os relacionamentos existentes no tesauro que estava sendo criado ou em manutenção.

A criação de tesauros pressupõe fases já estabelecidas nos processos realizados pelos profissionais de informação, envolvendo atividades de seleção de documentos representativos, a extração dos termos, a geração dos conceitos e o estabelecimento da estrutura, ou seja, a criação dos relacionamentos entre os termos (BITI, 2005; CAMPOS, 2001; CINTRA et. al; 2002; DODEBEI, 2002; GOMES e CAMPOS, 2004). Estes quatro passos básicos orientaram a arquitetura de construção do AlfaThes, apresentado na Figura 7.

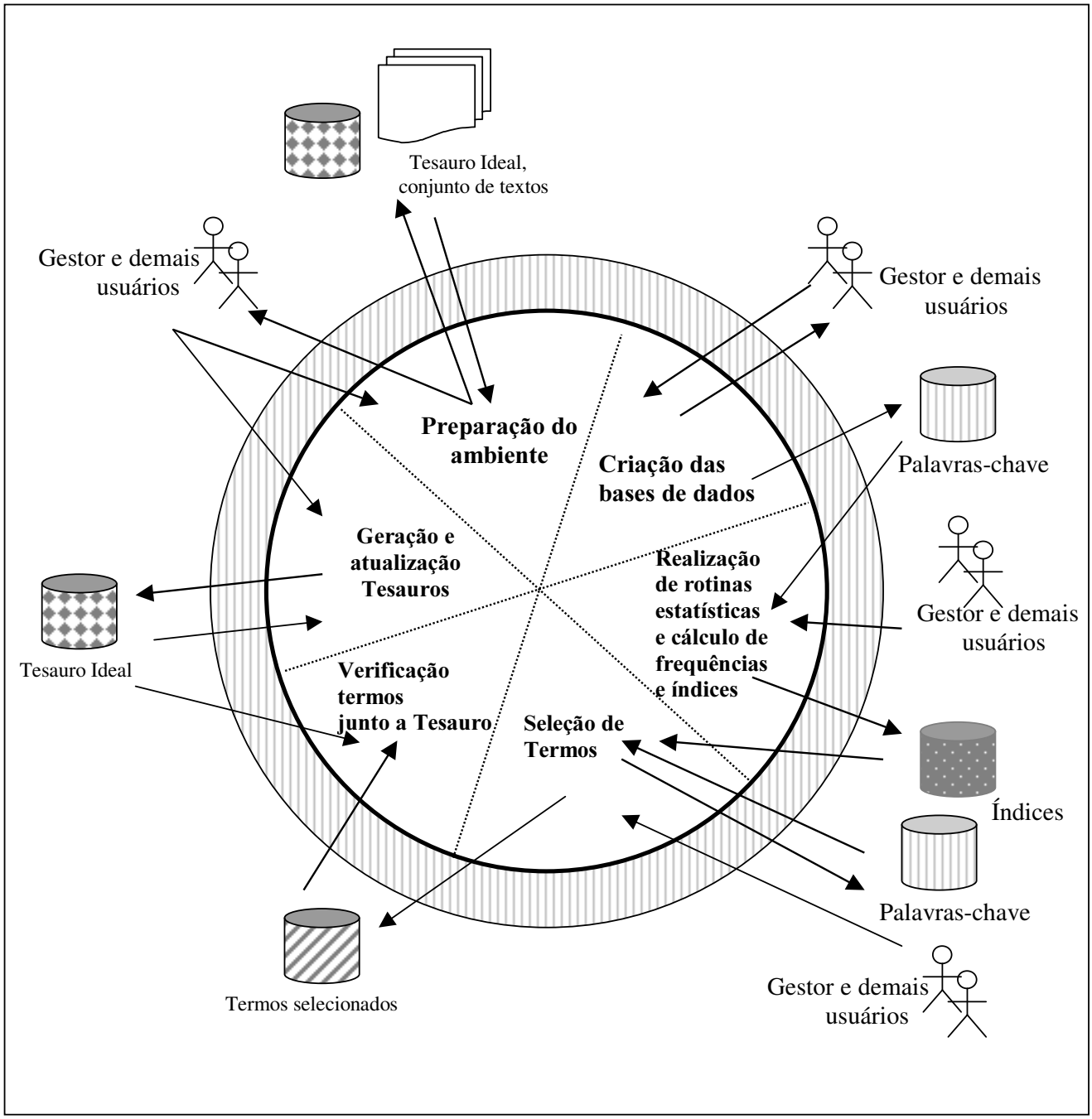

Figura 7 - AlfaThes - Ambiente de geração e manutenção semi-automática de tesauros

O AlfaThes foi construído em seis etapas, descritas a seguir:

- Preparação do ambiente e criação das bases de dados

- Realização de rotinas estatísticas
- Cálculo de freqüências e índices
- Análise e seleção de termos
- Verificação de termos em tesauro ideal
- Geração e atualização do tesauro

A construção de tesauros prevê a fase de planejamento, quando são estabelecidas as políticas que o nortearão. Antecedendo a isto, todo tesauro tem uma motivação centrada em determinado assunto, em uma área do conhecimento, na missão institucional, dentre outras.

A Ciência da Informação foi escolhida como área do conhecimento para o AlfaThes. Como políticas, decidiu-se que os termos deveriam retratar esta área em sua forma mais ampla, ou seja, os termos deveriam ser selecionados levando em conta sua interdisciplinaridade com a Biblioteconomia, Arquivologia, Museologia, Ciência da Computação, Economia, Comunicação, Ciências Cognitivas, Administração, Ciências Políticas, História, Filosofia, Direito, Linguística, Sociologia, Educação, entre outras. E seriam, sempre que possível, tomados no singular e em gênero masculino; que as siglas seriam consideradas somente em casos de alto grau de importância e que deveriam ter seu nome explicitado em notas de escopo e como não-descritores; que os nomes de pessoas e de organizações não seriam incluídos; que termos adicionais, não existentes entre as palavras-chave dos documentos selecionados, poderiam fazer parte do tesauro, desde que fossem consideradas importantes pela equipe do projeto.

3.2.1 Preparação do ambiente e criação das bases de dados

Esta fase correspondeu aos processos de preparação dos dados para execução das rotinas previstas no AlfaThes e envolveu os seguintes passos:

3.2.1.1 Preparação da entrada de dados: seleção dos documentos

Os tipos de documentos escolhidos para o ambiente foram os textos científicos: artigos, recensões, comunicações, relatos de experiência e artigos de revisões de literatura. Esta escolha baseou-se no fato de ser possível identificar nesses tipos de texto as unidades eleitas para estudo: o título, o resumo, o texto e a bibliografia, apoiado nos trabalhos de França et al.(2004), Cunha (2004) e nas normas técnicas NBR10520, NBR6023, NBR6022, NBR6024, NBR6028, da ABNT.

Foram selecionados para esta edição do AlfaThes textos científicos em formato eletrônico, disponibilizados pelas revistas *Datagrama Zero* e *Ciência da Informação* disponíveis nos respectivos endereços eletrônicos, em janeiro de 2005. A escolha deu-se por diversos motivos:

- ter sido possível reunir textos de duas coleções de um mesmo período (1999-2004);
- uniformidade dos temas tratados nos dois periódicos, envolvendo textos relacionados à Ciência da Informação;
- disponibilidade dos textos em formato digital, em sua forma integral e na Web;

- vínculo com as garantias literária, de uso e estrutural, propostas como suporte teórico a este trabalho;
- regras de submissão necessárias à garantia que vem das unidades de texto[20].

Observa-se ainda que o período escolhido foi direcionado primeiramente pela disponibilidade da revista Datagrama Zero, mas que corresponde a um período privilegiado na produção em Ciência da Informação, envolvendo temas relacionados a objetos de seu estudo e com suas áreas limítrofes. São exemplos os avanços na tecnologia da informação, nas ciências gerencias, na área da cognição, entre outros tantos.

A revista Datagrama Zero é um periódico eletrônico, bimestral, produzido pelo Instituto de Adaptação e Inserção na Sociedade da Informação (IASI), uma organização não-governamental, sem vinculação político-partidária ou religiosa, fundada em novembro de 1998 e dedicada a estudos e pesquisas sobre a Sociedade da Informação.

Nela, os textos estão reunidos por afinidade temática, englobando assuntos relacionados às áreas interdisciplinares da Ciência da Informação, como: Informação e Sociedade, Informação e Políticas Públicas, Informação e Filosofia ou Informação e Comunicação, sendo disponibilizados em versão integral no formato *html*.

O periódico possui regras de submissão de artigos disponibilizadas no próprio *site.* Destas, destacam-se: a regra referente à língua em que deverão estar os artigos: português ou espanhol; a exigência de um resumo temático, com cerca

[20] Ver o item 2.4.2

de cem palavras, em português (ou espanhol) e inglês; titulo em português (ou espanhol) e inglês e a exigência de um conjunto mínimo de cinco palavras-chave (DATAGRAMA ZERO, 2004). A primeira regra justifica a escolha da massa de teste, formada com todos os textos na língua portuguesa, o que resultou no total de 128 textos, no período de dezembro de 1999 a outubro de 2004; a segunda exige o resumo em todos os artigos e a terceira é válida para o objetivo do experimento, pela presença das palavras-chave.

A Revista Ciência da Informação é de responsabilidade do Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT). Seu primeiro número data de 1972. Teve periodicidade semestral até 1991, passando então a ser quadrimestral. É um periódico de atuação marcante no campo da ciência da informação e do setor de informação em Ciência e Tecnologia (C&T), publicando textos de especialistas nacionais e estrangeiros. A revista tem seus artigos indexados ou resumidos em *Paschal Thema: Science de L'Information, Documentation; Library and Information Science Abstracts; PAIS Foreign Language Index; Information Science Abstracts; Library Literature; Páginas de Contenido: Ciencias de la Información; EDUCACCION: Notícias de Educación, Ciencia y Cultura Iberoamericanas; Referativnyi Zhurnal: Informatika*.

Seu conteúdo traz artigos, recensões, entrevistas, comunicações, relatos de experiência, artigos de revisão de literatura e a seções *in memoriam*, ponto de vista e editorial. A revista publica trabalhos em português, espanhol, inglês e francês. Estão disponibilizados na Web os artigos desde o primeiro número de 1995. A revista possui normas editoriais, dentre as quais destacam-se: o título do trabalho deve *"ser breve e suficientemente específico e descritivo"*; o resumo, conter cerca de 200 palavras e estar de acordo com a NBR

6028/2003. As demais unidades de texto – numeração, bibliografia, citações – deverão estar de acordo com as normas NBR 6024/2003, NBR 6023/2002, NBR 10.520/2002 respectivamente.

A importância desta revista para a área de Ciência de Informação e seu rigor editorial justificaram sua escolha. Dela foram colhidos 186 textos em português, de diversos tipos, editados entre janeiro de 1999 e setembro de 2004.

Foi criado um diretório para cada coleção de artigos. No contexto deste trabalho, entende-se como diretório uma pasta de arquivos em computação e por coleção um conjunto de documentos colhidos de uma mesma fonte. O projeto contou então com duas coleções: Datagrama Zero e Ciência da Informação.

Todos os textos colhidos, originalmente em formato *html*, foram convertidos para o formato texto (*txt*) e o nome externo do arquivo seguiu a padronização: artigo + ano + mês + número do artigo.txt. Assim, o arquivo artigo1999jan1.txt refere-se ao texto de número um da edição de janeiro de 1999. Optou-se por não colocar o nome da coleção no nome externo do arquivo, já que os textos ficaram mantidos em diretórios distintos por coleção.

Foram considerados como atributos qualificadores de cada coleção a fonte de onde foram retirados os documentos e o período de produção dos mesmos, já que estes facilitariam, posteriormente, a recuperação de informações estatísticas por estes cortes.

3.2.1.2 A montagem de um tesauro ideal em Ciência da Informação

Na busca de um tesauro em língua portuguesa, no contexto da Ciência da Informação, constatou-se uma lacuna, em relação à atualidade. No Brasil, o tesauro especializado nessa área é o tesauro do IBICT, produzido em 1989, defasado portanto frente à velocidade de atualização desta área.

A necessidade tornou-se um desafio. Era preciso gerar um tesauro em Ciência da Informação, aqui considerado tesauro ideal, capaz de satisfazer as necessidades de processamento requeridas por este trabalho e de retratar as diversas áreas de domínio da Ciência da Informação; e que fosse construído levando-se em conta as garantias literária, de uso e estrutural. Para facilitar o entendimento do texto, a partir deste instante o tesauro ideal em Ciência da Informação, construído neste trabalho será chamado de TCI (Tesauro em Ciência da Informação).

A Ciência da Informação é um campo emergente no âmbito das Ciências Sociais que se caracteriza por estudos de cunho multidisciplinar e interdisciplinar. Como resultado, a sua representação temática oscila entre as especificidades do campo e as hibridações conceituais realizadas nas fronteiras do seu domínio. A geração de um tesauro de tipo ideal para a Ciência da Informação, como o TCI, deve levar em consideração essa particularidade, sob o risco de estabelecer uma estrutura pouco consistente e um facetamento que não reflita o seu real entrelaçamento conceitual, preocupação presente em todo o processo.

A escolha dos tesauros que iriam servir de base à criação do TCI – em português e em outras línguas - foi orientada pela legitimidade dos órgãos

responsáveis por sua criação, por seu uso consolidado na comunidade usuária e pela disponibilidade de acesso aos mesmos. Foram então eleitos quatro tesauros: o *Thesaurus of Information Science and Librarianship,* da *American Society for Information Science* (ASIS), o *Tesauro de Ciencias de la Documentación* - tesauro DOCUTES - da Universidade de León (Espanha), o *Tesauro en Biblioteconomía y Documentación* do *Centro de Información y Documentación Científica* (CINDOC) e o Tesauro em Ciência da Informação, do IBICT.

Os tesauros CINDOC e DOCUTES foram retirados da Web em seus respectivos endereços. O tesauro ASIS e o IBICT encontram-se disponibilizados em bibliotecas. A opção por mais de um tesauro nesta construção justifica-se no fato de que a Ciência da Informação conjuga outras ciências em seu corpo teórico e por se desejar buscar culturas e visões diferentes de representação de seus campos. A escolha do tesauro em língua inglesa deve-se à legitimidade do mesmo na comunidade mundial de Ciência da Informação. Os tesauros espanhóis foram eleitos por retratarem as áreas de Arquivologia, Documentação e Museologia, tendo como foco o registro documental.

O tesauro ASIS, em segunda edição, segue basicamente o proposto na edição anterior. A política de sua criação elegeu como objetivo uma cobertura mais profunda nos campos da Ciência da Informação e Biblioteconomia, para a indexação nestas áreas, sem esquecer os campos relacionados e periféricos mais fortemente relacionados, entre eles: Ciência da Computação, Lingüística e ciências relacionadas à cognição e ao comportamento. Outras áreas

periféricas foram incluídas, com uma cobertura limitada: Economia, Administração, Estatística e Sociologia (MILSTEAD, 1998).

O tesauro ASIS descreve e incorpora o caráter interdisciplinar da Ciência da Informação e suas interfaces mais evidentes, mas isto é feito de forma mais abrangente e horizontal, ficando o desejo de maior aprofundamento em alguns pontos. As manifestações da informação são ordenadas em disciplinas distintas, destacando-se as contribuições dos campos disciplinares à Ciência da Informação e sinalizando o aspecto informacional a ser organizado em cada campo (MOREIRA e MOURA, 2005).

O tesauro encontra-se delimitado a partir de assuntos tópicos. De acordo com a política de seleção dos termos, não foram incluídos nomes próprios de organizações, pessoas, programas, etc., com algumas exceções, como por exemplo, os nomes de ferramentas, tendo em vista sua representatividade na prática da indexação e catalogação. Outra exceção aberta diz respeito a termos fundamentais no contexto da Web, já que por sua disseminação e uso comum podem ser considerados nomes próprios (MILSTEAD, 1998).

Esta edição do tesauro ASIS incluiu 1353 descritores e 778 não-descritores, distribuídos em trinta e seis facetas. São elas: atividades e operações; negócio e gerência de operações; atividade de comunicação; operação de computador; atividade educacional; atividade geral; informação e operações em biblioteca; atividade socioeconômica; operação técnica e de produção; edifício e instalações; mídia de comunicação; tipos de documentos; tipos de documentos por disponibilidade, acesso, organização; tipos de documentos por conteúdo de informação, propósito; tipos de documentos por meio físico; campos e

disciplinas; equipamento, hardware e sistemas; conhecimento, informação, etc.; conhecimento, informação (por conteúdo); representação da informação; conhecimento e dispositivos de organização da informação; linguagem; elementos lingüísticos; funções naturais e eventos; redes; organizações; pessoa e grupos informais; mídia física; provedores de produtos e serviços; qualidade; qualidade em geral; qualidade humana; qualidade da informação e de dados; qualidade de sistema e equipamentos; pesquisa e métodos analíticos; aspectos sócio-culturais.

A equipe de montagem do tesauro vivenciou, entre as dificuldades enfrentadas para sua atualização, o dinamismo existente na Ciência de Informação, cujo campo de ação sofreu a velocidade de mudanças, principalmente após a Web, com a incorporação de novos termos e a eliminação de muitos outros. Foi necessário equilíbrio para incluir termos atuais, ou mesmo um jargão que poderia ser superado antes do tesauro ser publicado. Com base nessa premissa, a ausência de alguns termos foi justificada. A versão data de 1998 e alguns termos hoje conhecidos por todos, como XML, por exemplo, não constam dessa edição.

O *Tesauro de Biblioteconomía e Documentación* do CINDOC (CINDOC, 2005) foi proposto para ser uma linguagem controlada para a análise de conteúdo e recuperação de documentos incluídos na *Base de Datos de Biblioteconomía, Documentación y Política Científica* – ISOC-DC, produzida pelo CINDOC, do *Consejo Superior de Investigaciones Científicas* (CSIC) desde 1975. Sua criação teve como alvo suprir a falta de léxicos documentais em espanhol abrangendo os campos semânticos representados nos textos técnico-

científicos publicados na Espanha, muitos dos quais estão armazenados na mencionada base de dados.

A implantação deste tesauro possibilitou indexação mais homogênea dos documentos incorporados na base ISOC-DC, facilitando uma recuperação fácil e exaustiva, eliminando ambigüidades e dando uma visão da afinidade semântica entre os termos distintos, enriquecendo o trabalho dos documentalistas e ampliando o campo de busca do usuário. O tesauro proporciona à comunidade científica um conjunto estruturado de termos sobre a base de um sistema de conceitos para a organização do conhecimento biblioteconômico (CINDOC, 2005).

A ISOC-DC é uma base de dados bibliográficos de artigos de revistas espanholas sobre Biblioteconomia, Documentação, Arquivologia, Política Científica e Política de Informação, assim como artigos de congressos espanhóis nessas áreas. Permite a recuperação de informação por diversos critérios de busca. A temática abrange as áreas de Antropologia, Arqueologia e Pré-História, Belas Artes, Biblioteconomia e Documentação, Direito, Economia, Educação, Filosofia, Geografia, História, Lingüística, Literatura, Psicologia, Ciências Políticas e Sociologia, Urbanismo e Estudos americanistas. A vastidão do universo para o qual foi criada foi um dos motivos de sua escolha como um dos tesauros para a composição do TCI.

O tesauro, orientado para sua disponibilidade na Web (DOCUTES, 2005), foi elaborado pelo setor de *Biblioteconomía y Documentación* da *Universidad de León*, através de um projeto de investigação científica patrocinado pela *Junta de Castilla* no ano 2000 e faz parte de um objetivo maior de potencializar

práticas e ações na área através de tutoriais desenvolvidos a partir de material adequado. É utilizado nas atividades docentes da universidade e, por esta razão, tem disponibilidade constante na Web, onde é possível encontrar a relação completa dos termos distribuídos nas seguintes facetas: Ciência da documentação: Historia. Teorias. Sistemas; Informação. Documentos. Fontes de informação; Investigação e metodologia documental; Profissionais e usuários; Representação e recuperação de informação; Sistemas de informação e Tecnologias da informação.

Para a construção do tesauro a equipe do projeto DOCUTES utilizou-se do *MultiTes* (MULTSYSTEMS, 2005), software específico para auxílio na criação e edição de tesauros eletrônicos.

O Tesauro de Ciência da Informação do IBICT (IBICT, 1989) foi desenvolvido com vistas a atender às necessidades de indexação dos documentos existentes no acervo desse instituto e está orientado por sete grandes categorias, aqui entendidas como facetas: Informação; Documento; Unidade de informação; Planejamento; Processos e serviços de informação; Transferência e uso da informação e Profissão.

Observou-se uma articulação de suas facetas em torno do conceito de informação; mas não reflete o avanço do campo e de suas interfaces disciplinares. Há um número excessivo de termos que demarcam um atraso em relação à nova configuração do campo. Este fato era esperado já que o tesauro data de 1989 e, desde então, foram muitas as atualizações existentes na Biblioteconomia e na Ciência da Informação.

A edição de 1989 traz toda documentação do processo de sua criação, explicitando as políticas e escolhas feitas na ocasião. Já esteve disponibilizado na Web, mas no instante da coleta dos dados não nos foi possível encontrá-lo nesta mídia.

3.2.1.3 A metodologia empregada

Para a construção do TCI alguns passos foram previamente estabelecidos, em conformidade com os autores BITI (2005), Campos (2001), Cintra et. al (2002), Dodebei (2002), Gomes e Campos (2004). Acredita-se que estes passos possam facilitar a construção de futuros tesauros em outras áreas, a partir de tesauros já existentes.

- **Estabelecimento dos objetivos**: O objetivo do tesauro foi definido em três pontos: primeiramente, ele deveria servir de apoio a este experimento de geração semi-automática de tesauros; em segundo lugar, deveria fornecer subsídios para indexação de artigos científicos em língua portuguesa e em terceiro lugar, deveria estar disponível na Web para futuras pesquisas utilizando tesauros e este ambiente.

O campo temático foi delimitado pelas áreas de Arquivologia, Biblioteconomia, Ciência da Informação e Documentação. Estas áreas do conhecimento foram tomadas a partir de seu próprio foco e das áreas que compõem o universo de sua produção. Primeiramente elegeu-se a área de Ciência da Informação como central para o propósito do tesauro a ser construído. A seguir, analisou-se sua relação com a área de Biblioteconomia, com levantamento das áreas limítrofes mais próximas dos conteúdos destas duas áreas ou de outras áreas cujo

relacionamento com a Ciência da Informação e com a Biblioteconomia revela interdisciplinaridade. São elas: Administração, Arquivologia, Museologia, Ciências Cognitivas, Ciências da Computação, Economia, Educação, Linguística e Sociologia. Estas áreas compõem o círculo primeiro de relacionamento com a Ciência da Informação e a Biblioteconomia. Seguindo a elas, estão as áreas de Filosofia, Ciências Políticas, História e Direito, por suas relações com as áreas centrais os com as áreas do primeiro círculo.

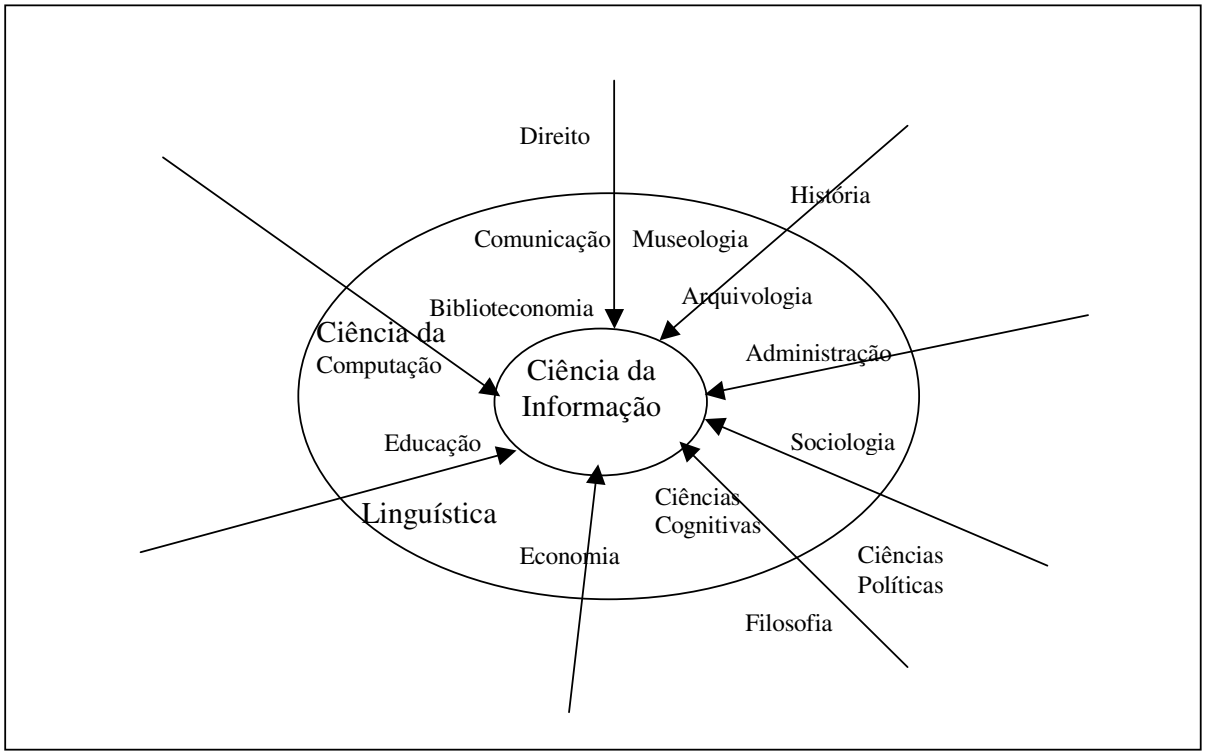

Figura 8 - Áreas limítrofes à Ciência da Informação

A Figura 8 ilustra o relacionamento estreito entre estas áreas e aponta para as seguintes questões:

- a alfabetização informacional, através dos conteúdos da Educação, da Linguística e da Ciência da informação;

- a comunicação mediada por computador, através das demandas da comunicação e das soluções proposta pela Ciência da Computação;
- a educação em Ciência da Informação, como peculiaridades da Educação no domínio específico da Ciência da Informação;
- a liberdade intelectual, através do Direito, da Comunicação e da Sociologia;
- as questões relativas à Informação e Sociedade, pelos conteúdos de Sociologia e Ciência da Informação e Ciências Políticas;
- a automação dos processos de indexação e o processamento de linguagem natural através da Linguística, da Biblioteconomia e da Ciência da Computação;
- os pontos relativos à gerência de informação, de recursos informacionais, de biblioteca, através dos conteúdos da Administração, da Ciência da Informação e da Biblioteconomia.

- **Definição da equipe**: Partiu-se do princípio de que a equipe de construção de um tesauro deve ser representativa da sua área. Neste caso específico, as atividades foram realizadas pelo autor e tiveram a supervisão da professora orientadora. É importante observar que a construção de tesauros deve sempre prever um trabalho em equipe, constituída sempre que possível por pessoas com conhecimento do campo temático, conhecimento de elaboração de tesauros e abertura para novos horizontes.

Uma aluna do curso de Biblioteconomia ECI/UFMG auxiliou a digitação dos termos do tesauro construído para testes. Em muitas reuniões de debate sobre pertinência de termos e sobre o tratamento de palavras, ela também esteve presente e de forma participativa.

Um aluno do curso de Ciência da Computação PUCMinas auxiliou nas atividades referentes à programação, sob supervisão do autor, tendo participação ativa na construção física dos bancos de dados gerados e do modelo do ambiente.

A equipe formada para o ambiente AlfaThes contou então com conhecimentos oriundos da Ciência da Informação e da Ciência da Computação. Esse era o interesse desde o início do projeto: vivenciar esta interdisciplinaridade, permitindo que a aproximação de campos distintos do saber na solução de problemas a elas pertinentes levasse os envolvidos no processo a compartilhar não só conhecimento, mas também metodologias de trabalho que atendam às áreas envolvidas na solução, conforme aponta Domingues (2005).

- **Eleição de tesauros já existentes na área**: Esta atividade foi orientada para os objetivos de criação do tesauro e para o público que dele fará uso. Estabeleceu-se, previamente, que os tesauros eleitos deveriam ser ferramentas consolidadas pela comunidade da área de Ciência da Informação, possuir filosofia e estruturação próximas às desejadas, e que deveriam estar alinhados ao suporte teórico deste trabalho: as garantias literária, de uso e estrutural. No caso do TCI, a legitimidade na comunidade, o grau de utilização e a disponibilidade dos tesauros foram os norteadores, resultando na escolha dos tesauros da ASIS, DOCUTES, CINDOC e o do IBICT.

- **Coleta dos dados**: Os dados disponíveis em meio eletrônico foram colhidos diretamente da Web. Os provenientes dos tesauros ASIS e IBICT foram digitados a partir das respectivas edições impressas. Foram criados arquivos diferentes para cada um dos tesauros, respeitando suas facetas, hierarquias e os relacionamentos entre os termos Foi importante eleger um deles como base para o novo tesauro. Foi escolhido o tesauro ASIS como ponto fundamental de partida para comparações e efetivação de seus próprios termos e dos termos e relacionamentos dos demais tesauros.

A Tabela 1 mostra o número inicial de termos coletados de cada tesauro. Observa-se que alguns tesauros possuíam número maior de termos do que aqueles relacionados neste trabalho, uma vez que alguns termos foram descartados durante o processo de tradução, por não pertencerem ao universo da língua portuguesa ou serem termos específicos dos países e das culturas do tesauro original. Os termos foram separados mantendo suas hierarquias originais, conservadas à parte para a redistribuição dos mesmos.

Tabela 1 - Total de termos retirados dos tesauros

Tesauro de origem	Total de termos retirados
IBICT	947
CINDOC	1145
DOCUTES	1515
ASIS	1623
Total de termos	5230

- **Estabelecimento de políticas**: Nesta etapa foram previstas as regras e normas a serem seguidas na elaboração do tesauro. Estabeleceu-se como

padronização para os termos o uso da forma singular; evitou-se a utilização de termos emprestados de outras línguas com exceção daqueles que, de acordo com o senso comum, são utilizados no Brasil; foram evitados nomes próprios e de organizações; nos casos em que as siglas foram incluídas, optou-se por criar novo termo como não-descritor com o nome completo; parênteses foram utilizados para incorporar um qualificador ao descritor, mas de forma geral esta prática foi evitada.

- **Tradução dos tesauros de origem**: Como três dos tesauros escolhidos não se encontravam em língua portuguesa, foi necessária uma fase de tradução dos mesmos, que ficou sob a responsabilidade deste autor, com a revisão feita pela professora orientadora. Essa tradução procurou avançar além da tradução de palavras, buscando sempre termos empregados na cultura do tesauro a ser construído e eliminando termos e relacionamentos que não agregavam valor para indexação na língua portuguesa.

Observou-se que a tarefa de tradução realizada por apenas uma pessoa colabora para conduzir à utilização de termos semelhantes. Sendo feita por uma só pessoa, e diferente do tradutor, a revisão favorece a escolha de um termo entre os sinônimos e facilita o encontro de termos incongruentes ao propósito da área. É interessante que a revisão da tradução seja feita por profissional que conheça a área temática do tesauro. Os termos precisam retratar a cultura da língua para a qual estão sendo levados e, principalmente, o emprego e uso dos mesmos pela comunidade científica e de usuários, respectivamente.

- **Escolha das facetas do futuro tesauro**: Nesta fase, a partir das facetas encontradas nos tesauros existentes e em conformidade com os objetivos do tesauro proposto, elegeram-se as facetas que iriam compor o futuro tesauro. Foi observado que esta não é uma etapa definitiva e durante as operações de junção dos termos e de estruturação do tesauro, descritas nos itens seguintes, foi necessário retomar este passo e rever as facetas eleitas. As revisões realizadas proporcionaram novos arranjos dos termos nas facetas escolhidas.

Para o novo tesauro, foram selecionadas facetas dos quatro tesauros que dele foram fontes. Em alguns casos a faceta herdou o nome original da fonte utilizada. A Tabela 2 mostra as facetas criadas no TCI e a fonte dos termos e relacionamentos para cada uma delas.

Tabela 2 - Fontes das facetas criadas

Faceta	**Tesauro fonte**
Arquivologia	CINDOC
Campos e disciplinas	ASIS, CINDOC, DOCUTES
Ciência da Informação	IBICT
Conhecimento e informação	ASIS
Documentação	ASIS, CINDOC, DOCUTES, IBICT
Fontes de Informação	DOCUTES
Informação e operações em bibliotecas	ASIS, CINDOC, DOCUTES, IBICT
Mídia Física e de comunicação	ASIS
Museologia	CINDOC
Organizações	ASIS
Pesquisa e métodos analíticos	ASIS
Pessoas, profissionais, grupos informais	ASIS, CINDOC, DOCUTES, IBICT
Tecnologia da Informação	ASIS, CINDOC, DOCUTES
Unidades de Informação	ASIS, CINDOC, IBICT

- **Junção e comparação entre os termos**: A atividade prevê a reunião dos termos existentes nos tesauros organizados nas facetas escolhidas anteriormente. Utilizou-se de recursos de computação, como planilhas e gerenciadores de bancos de dados, para efetivação da comparação do emprego dos termos e geração do novo tesauro. A atividade englobou os seguintes passos:

    - reunião dos termos dos tesauros existentes nas facetas escolhidas conservando a relação hierárquica previamente existente;
    - análise, remanejamento e/ou eliminação de termos coincidentes e arranjo da estrutura hierárquica;
    - transformação de descritores em não-descritores, e vice-versa, conforme os objetivos do tesauro;
    - revisão da nova estrutura hierárquica.

A utilização de *softwares* para auxílio à estruturação de tesauros facilita este processo. No caso do TCI, os termos foram dispostos em sua relação hierárquica em documentos textos antes de serem transpostos para outros *softwares*. Planilhas eletrônicas e gerenciadores de bancos de dados foram utilizados para verificação de termos coincidentes e de semelhanças de grafias.

O emprego destes instrumentos foi fundamental para o auxílio no tratamento das palavras traduzidas, principalmente na eliminação de termos que não fazem parte da cultura brasileira, da sinonímia e da homonímia. Os termos foram inicialmente introduzidos nas facetas escolhidas e, através da utilização de listagens, os termos receberam os devidos tratamentos. A tradução facilitou

a verificação de termos com a mesma grafia, resultando na possibilidade de sua eliminação e de alocação dos termos nas facetas.

A versão acadêmica do *Thesaurus Construction System version 8* (TCS-8) (WEBCHOIR, 2005) foi utilizada para a construção do tesauro a partir dos termos e das relações hierárquicas já estabelecidas. O TCS-8 é um sistema para auxílio à construção de tesauros que apresenta flexibilidade, robustez e facilidades de manuseio, permitindo aos usuários criar tesauros e outros vocabulários hierárquicos através de ferramentas para verificação de duplicidade e documentação dos termos, para a construção de hierarquias e relacionamentos, inclusive distinguindo os tipos relacionamentos. Orienta-se pela *National Information Standards Organization* (NISO) padrão z39.19, permitindo que o usuário, caso deseje, crie suas próprias regras de estrutura.

O emprego de softwares de construção de tesauros foi importante nesta etapa pela facilidade de visualização das estruturas hierárquicas, para construção de notas de escopo e demais itens de documentação, além de agilizar o estabelecimento de relações entre os termos e sua categorização.

Na versão inicial do TCI, os 5230 termos resultaram em 2230 termos, sendo que 292 destes eram não-descritores. Sua base maior permaneceu com o tesauro ASIS, sendo que também deste alguns termos foram eliminados tendo em vista sua pouca utilização no Brasil. Outros foram acrescentados, e novos relacionamentos foram criados. Dos tesauros DOCUTES e CINDOC foram selecionados, principalmente, os termos e os relacionamentos representativos da área de documentação e arquivologia.

Os termos foram introduzidos no software TCS-8 a partir das facetas escolhidas. A primeira análise do produto revelou que o mesmo possuía termos que necessitavam de uma redistribuição, assim como muitos outros deveriam ser utilizados como não-descritores. Seguiram-se atividades de documentação dos termos (notas de escopo) e estabelecimento dos relacionamentos, assim como eliminação de termos e reorganização destes como descritores e não-descritores.

- **Estruturação do novo tesauro**: As garantias literária, de uso e estrutural tornaram-se também o norte do TCI. Cada termo analisado era revisto tendo em vista sua utilização na área. Alguns termos não conhecidos pela equipe passaram por processo de consultas a bibliografias ou informações na Web. Embora já estivesse estabelecida pelos tesauros de origem, a garantia literária precisava ser preservada, justificando estas consultas.

A organização dos conceitos foi conduzida pelo processo de indução/dedução. À medida que as relações foram sendo estruturadas, novos conceitos foram acrescentados, complementando a organização de cada faceta do futuro tesauro. De forma genérica, três tipos de relacionamentos direcionam a elaboração de tesauros: a) relacionamento de equivalência; b) relacionamento hierárquico; c) relacionamento associativo.

Cada um deles possui sua propriedade de reciprocidade. O relacionamento associativo e o de equivalência são simétricos enquanto o hierárquico é assimétrico; o que equivale a dizer que, dados dois termos A e B, se A está relacionado a B por uma associação qualquer, B também está relacionado a A pela mesma associação. Por sua vez, se A e B são termos equivalentes, um

destes termos deverá ser escolhido como descritor. O outro termo fará parte do tesauro como não-descritor através da orientação USE. Observa-se que se trata de um relacionamento simétrico com o correspondente inverso USADO PARA. Nem sempre a relação de equivalência diz respeito à sinonímia em um tesauro, como pode ser observado no emprego de políticas como a adotada para o uso de siglas para as quais os nomes completos correspondem a elas. Muitas vezes termos não sinônimos são empregados por usuários distintos com o mesmo significante e representados no tesauro a partir destas relações.

No TCI, quando se fez necessária a inclusão de siglas, foram também incluídos os termos não-descritores com seu significado. Para termos não sinônimos utilizados com o mesmo significante ou associados de alguma forma foram introduzidos relacionamentos no TCI.

O relacionamento hierárquico é uma característica básica dos tesauros distinguindo-os das demais listas de termos, glossários ou outro tipo de vocabulário controlado. Segue as regras gerais de uma hierarquia, onde um termo pode ter diversos outros a ele subordinados mas subordina-se apenas a um único termo. Contudo, podem ocorrer casos de polihierarquia nos quais um termo poderá pertencer a facetas distintas, o que não é permitido pelo TCS-8. Quando ocorreram estes casos, utilizou-se do artifício de adicionar ao termo um asterisco (*) diferenciando-o do termo originalmente incluído. Em alguns casos, relações associativas entre esses termos foram criadas, no TCI, por exemplo, o termo ‘Almanaque’, pertencente ao termo genérico ‘Obra de referência’ da faceta ‘Documentação’ e o termo ‘Almanaque*’, pertencente ao termo genérico ‘Fonte primária’ da faceta ‘Fontes de informação’. Um relacionamento associativo foi criado entre esses termos.

Ainda quanto às hierarquias, no contexto dos tesauros, um descritor dito superordenado representa uma classe ou um todo, e os descritores a ele subordinados são referidos como seus membros ou partes.

Alguns autores, entre eles BITI (2005), classificam as relações como lógicas e ontológicas. No primeiro grupo encontram-se as relações hierárquicas, sendo que as de equivalência e as partitivas (*é parte de*) compõem as do segundo grupo. Para esses autores, uma relação partitiva não pode ser considerada hierárquica, pois em muitos casos as partes de um objeto possuem cada qual sua própria classe *(ex: biela, pistão, eixo de manivela, como partes de um motor)* e em alguns outros o objeto pode ser visto como parte e também como objeto independente *(ex: a Terra enquanto planeta é parte do Sistema solar, mas pode ser vista independentemente do sistema, por suas características próprias como coloração, dimensão, etc ...)*. Citam também casos onde o todo não se configura como um objeto mais importante ou determinante da existência da parte, como no caso, por exemplo, de carro (todo) em relação ao motor (parte/componente) onde existe uma necessidade da parte para funcionamento do todo (BITI, 2005). No caso do TCI estes pontos não foram considerados embora haja concordância com esses autores.

- **Revisão e atualização**: Esta etapa garante a longevidade de um tesauro. Nela estão contidas as atividades de manutenção e atualização dos termos e de suas relações. O presente trabalho prevê esta atividade de forma semi-automática. No caso do TCI, os termos gerados pelo ambiente AlfaThes poderiam ser utilizados para sua atualização. Optou-se por não fazê-lo antes de melhor avaliação do produto. O TCI estará disponibilizado na Web e prevê-

se a interação com os usuários da comunidade de Ciência da Informação para suas atualizações, que deverão ser trimestrais e os termos oriundos do AlfaThes farão parte de sua primeira atualização. O TCI também será utilizado em práticas docentes que possibilitarão sua atualização, permitindo futuros trabalhos.

O TCI foi concluído em outubro de 2005. Possuía então 1891 termos, divididos em 1694 descritores e 197 não-descritores, e será disponibilizado na Web. A Tabela 3 apresenta o número de descritores de cada faceta.

No que se refere às garantias, preocupação inicial de todo o processo, a literária e a de uso foram herdadas dos tesauros que serviram de base ao trabalho, tendo em vista o cuidado na tradução nesse sentido. O grande desafio foi a garantia estrutural já que aspectos ligados à cultura ficaram visíveis na relação de termos. Além disso, foram reunidos tesauros criados com objetivos distintos.

Tabela 3 - Número de descritores do TCI em cada faceta

Faceta	Número de descritores
Arquivologia	85
Informação e operações em biblioteca	330
Campos e disciplinas	205
Ciência da Informação	65
Conhecimento e informação	23
Documentação	158
Fontes de Informação	123
Mídia física e de comunicação	26
Museologia	43
Organizações	17
Pesquisa e métodos analíticos	80
Pessoas, profissionais e grupos formais	64
Tecnologia da Informação	358
Unidades de Informação	117

O Anexo 1 apresenta algumas páginas do TCI.

3.2.2 Criação das bases de dados

Esta fase correspondeu à criação das bases de dados de artigos e de palavras-chave, seguindo o modelo de entidades e relacionamentos (ELMASRI e NAVATHE, 2000) apresentado na Figura 9. Para maior facilidade no texto, a partir de instante, será utilizada a palavra 'artigo' como equivalente a todos os tipos de texto extraídos dos dois periódicos[21].

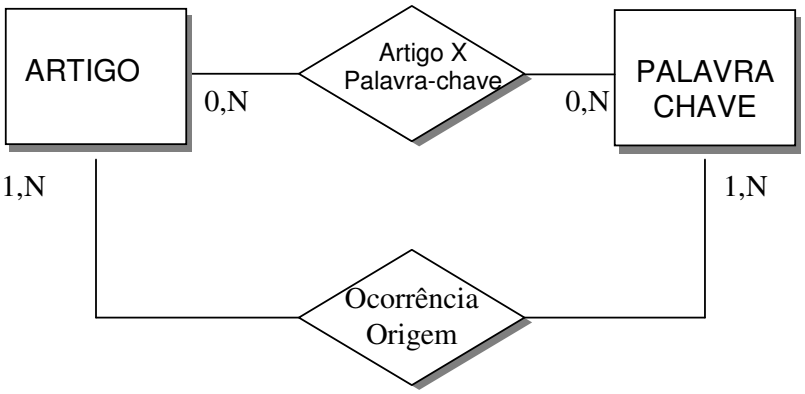

Figura 9 - Modelo de Entidade e Relacionamentos para as bases Artigo e Palavra-chave

A tabela Artigo contém as colunas Número do artigo (NumArt), Título do artigo (TitArt), Nome Externo do Arquivo (NomExtArq), Ano de publicação (AnoArt). A tabela Palavra-chave foi criada contendo as colunas código da palavra-chave (CodPalCha) e a palavra-chave (PalCha). A relação 'Ocorrência Origem' possui informações da freqüência de cada palavra-chave em seu artigo de origem, especificamente suas freqüências no título, no resumo, no texto e na bibliografia. Possui as colunas código do arquivo (CodArt), código da palavra-chave (CodPalCha), ocorrências no título (OcoTit), ocorrências no resumo (OcoRes), ocorrências no texto (OcoTex) e ocorrências na bibliografia

[21] Ver o item 3.2.1.1 deste trabalho.

(OcoBib). Entende-se por freqüência em estatística ao número de observações em uma classe não sobreposta (ANDERSON, et. al, 2002). Assim, esse conceito, neste trabalho, significa a ocorrência da palavra-chave em cada parte específica do texto.

A relação 'Artigo X Palavra-Chave' denota a ocorrência de uma mesma palavra-chave em mais de um artigo e possui as colunas Número do Artigo (NumArt) e Código da palavra-chave (CodPal), assim como os campos com a freqüência desta palavra em cada unidade do artigo da base, ou seja, no título (OcoTit), resumo (OcoRes), texto (OcoTex) e bibliografia (OcoBib) de cada artigo. A carga das freqüências nestas relações foi feita no procedimento 3.2.2.2 deste projeto, apresentado adiante.

3.2.2.1 Carga dos dados de artigo

Os artigos foram incluídos na base Artigo através de programação, numerados por ordem de entrada, sendo que foram lidos em ordem de classificação do nome externo. Esta operação equivale à função de documentação da base de documentos que compõe um tesauro (BITI, 2005). O título do artigo foi retirado através de expressão regular na programação. Na base de dados 'Artigo' foram gravados os atributos: número do artigo, título e ano de publicação, sendo que este último foi retirado do nome externo do arquivo. Para inserção dos artigos, o programa verificou se os mesmos possuíam as quatro unidades básicas de um texto, eleitas neste projeto: título, resumo, texto e bibliografia, além da obrigatoriedade de existência de palavras-chave. Naquele instante, quatorze textos apresentaram problemas pela inexistência de bibliografias, sendo cinco deles da coleção *Ciência da Informação* e nove da coleção *Datagrama Zero*. A

base de artigos passou a ter então 300 artigos, que constituíram a massa de dados deste projeto.

3.2.2.2 Criação da tabela palavras-chave

A carga da tabela Palavra-chave correspondeu à fase de extração das palavras-chave dos textos, para a qual foi criado um programa em PHP com as seguintes funções:

- Ler cada artigo de cada coleção, inserindo dados em Artigo;
- Para cada artigo, selecionar as palavras-chave, adicionando as mesmas na tabela Palavra-chave. Esta inserção é única para cada palavra e o programa verifica a sua existência antes de inseri-la. Caso já exista, o programa utiliza o código da palavra existente, mantendo a unicidade;
- Para as palavras-chave de cada artigo, o programa verifica a sua freqüência no artigo de origem em cada uma das quatro unidades do texto, e armazena os dados da tabela OcorrênciaOrigem, possibilitando visão das palavras-chave existentes em um texto específico, em uma coleção, no conjunto das coleções.

A programação utilizou funções de reconhecimento de padrão para verificação da freqüência de palavras-chave nos artigos de origem. Foram construídas expressões regulares para localização das palavras-chave, conforme o disposto em cada coleção, assim como também para o reconhecimento das unidades de texto integrantes do artigo científico.

A freqüência de palavras-chave está relacionada ao número de sua ocorrência nos textos selecionados, de acordo com Salton (1981), para quem freqüência é considerada tendo em vista o aparecimento da palavra no documento; de forma diferente desse autor, introduzimos o conceito de freqüência para unidades específicas de um texto científico (título, resumo, texto e bibliografia), analisando também a freqüência a partir do conceito de coleção e do conjunto de coleções.

Acredita-se que os subtítulos existentes em textos científicos e as citações representam também pontos de referência; mas, não foram objetos de análise deste trabalho.

Escolheu-se como grão para o estabelecimento da freqüência o texto e suas unidades, o que permitiu que o cálculo se estendesse para uma coleção específica e para o conjunto das coleções. Assim, foi construído um programa, em PHP, que lia cada texto da base 'Artigo', armazenando, na relação 'OcorrênciaOrigem', a freqüência de cada uma de suas palavras-chave em seu título, seu resumo, seu texto e sua bibliografia. Procedimento análogo foi feito para armazenar dados na relação 'Artigo X Palavra-Chave', agora verificando todas as palavras-chave coletadas com cada texto da base[22].

3.2.2.3 Tratamento das palavras-chave

Tendo em vista a experiência do gestor com as linguagens de indexação, considerou-se fundamental sua participação no processo de tratamento de palavras-chave , ao qual corresponderam os seguintes passos:

[22] Verificar detalhes do modelo de dados na Figura 9, p. 124.

- eliminação das ocorrências de plural-singular, masculino-feminino e de sinônimos no arquivo de palavras-chave;
- eliminação de palavras-chave que o gestor do projeto considerou não relevantes para o tesauro;
- substituição da palavra-chave por outra palavra mais comumente utilizada na área de Ciência da Informação.

Cabem algumas reflexões, pois as palavras-chave são indicadas pelos autores dos textos. No caso do AlfaThes, os textos são de cunho científico e sabe-se que os autores possuem domínio da área de conhecimento para a qual escrevem. E é relevante também observar que são pessoas interessadas na leitura e recuperação de seus textos, o que os leva a uma aproximação das palavras do universo disponível na área. Porém, nem sempre as palavras-chave escritas em linguagem natural expressam termos condizentes com uma linguagem de indexação, fato que reforça a necessidade do tratamento, realizado em duas etapas:

a) Criação da tabela de grupos de palavras

As palavras-chave encontradas nos artigos foram classificadas em grupos de áreas de conhecimento que possuíam interdisciplinaridade com a Ciência da Informação (Tabela 4). Alguns termos não possuíam característica específica de apenas uma área do conhecimento; outras vezes não foi possível identificar um grupo para a palavra-chave. Em ambos os casos, essas palavras foram classificadas em uma área denominada 'limítrofe'. Esta classificação permitiu o agrupamento das palavras para o devido tratamento. Ao gestor coube indicar o grupo ao qual a palavra-chave

pertenceria. Uma nova tabela foi inserida no banco de dados, a relação Grupo, e o modelo de entidades e relacionamentos do experimento passou a ser o apresentado na Figura 10.

Tabela 4 - Grupos ou grandes áreas do conhecimento

Sigla	Descrição
AD	Administração
AR	Arquivologia
BA	Belas Artes
BI	Biblioteconomia
CC	Ciência da Computação
CG	Ciência Cognitiva
CH	Ciências Humanas
CI	Ciência da Informação
CM	Ciências Médicas
CP	Ciências Políticas
CS	Ciências Sociais
EC	Economia
ED	Educação
FI	Filosofia
HI	História
L	Limítrofe
LI	Linguística
MS	Museologia
PI	Psicologia
SO	Sociologia

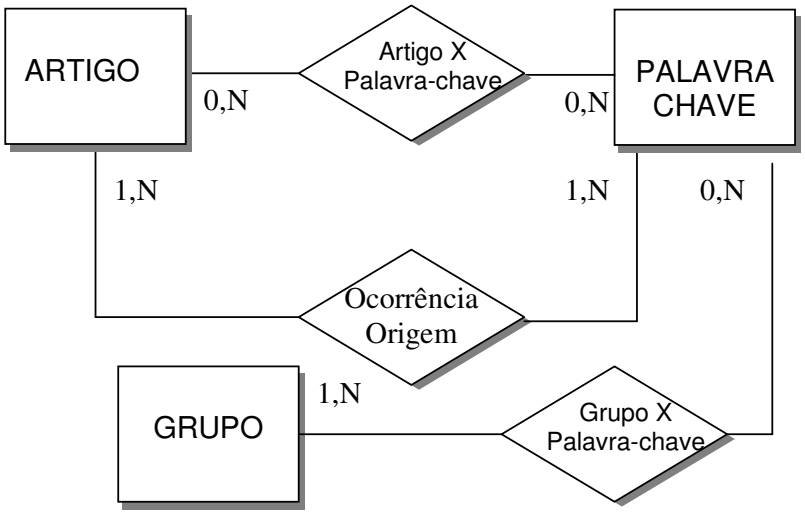

Figura 10 - Modelo Entidade e Relacionamentos com entidade grupo

A princípio, foi estudada a possibilidade de uma programação interativa onde o indexador informaria sobre esses grupos. Observou-se, entretanto, que ao gestor interessava a visão do conjunto dessas palavras para decisão do grupo ao qual pertenciam e por isso foi feita a opção pelo relatório, o que o deixou mais à vontade tendo as folhas do relatório espalhadas na mesa, com a possibilidade de reconsiderar sua decisão de colocar uma palavra em determinado grupo, tendo em vista o reconhecimento de outra palavra.

b) Tratamento de palavras-chave

Com as palavras-chave classificadas em grupos, foi emitido um relatório contendo o código da palavra, a palavra-chave, o título do artigo de onde a palavra havia sido retirada, o nome do arquivo externo do artigo e o grupo no qual a palavra fora classificada. Uma página exemplo deste relatório, classificado por grupo/palavra-chave para facilitar seu manuseio, é mostrada na Tabela 5.

Tabela 5 - Relatório para tratamento de palavras-chave

Código Palavra	Palavra	Código Artigo	Título Artigo	Arquivo	Grupo
454	Administração de empresas	132	Desafios da sociedade do conhecimento e gestão de pessoas em sistemas de informação	artigo 2003 mai 6.txt	AD
451	Administração de pessoal	132	Desafios da sociedade do conhecimento e gestão de pessoas em sistemas de informação	artigo 2003 mai 6.txt	AD
543	Administração hospitalar/ tendências	161	Sistema de informação: instrumento para tomada de decisão no exercício da gerência	artigo 2004 jan 6.txt	AD
485	Alinhamento estratégico	141	Metodologia para projeto de planejamento estratégico de informações alinhado ao planejamento estratégico: a experiência do Senac-PR	artigo 2003 set 14.txt	AD
88	Ambiente de qualidade	24	Ambiente de qualidade em uma biblioteca universitária: aplicação do 5S e de um estilo participativo de administração	artigo 1999 set 11.txt	AD
506	Apoio à decisão	148	Investigação de práticas anticompetitivas: um sistema de informação para apoio à interpretação de legislação por agências reguladoras	artigo 2003 set 8.txt	AD
292	Aprendizado organizacional	71	Inteligência organizacional: um referencial integrado	artigo 2001 mai 5.txt	AD
81	Aprendizagem organizacional	22	Aprendizagem organizacional e informação	artigo 1999 set 1.txt	AD
744	Call Center	225	Fonte de Informação Estratégica e Não-Estratégica	Artigo 2001 junho 05.txt	AD
418	Capacitação de recursos humanos	121	A sociedade da informação no Brasil: um ensaio sobre os desafios do Estado	artigo 2003 jan 4.txt	AD
610	Centros de Competência	182	Dado, Informação, Conhecimento e Competência	Artigo 1999 dezembro 01.txt	AD
564	Competência	167	Identificando competências informacionais	artigo 2004 set 12.txt	AD
564	Competência	182	Dado, Informação, Conhecimento e Competência	Artigo 1999 dezembro 01.txt	AD
869	Competências	266	Competências no discurso oficial da formação de professores no Brasil [1]	Artigo 2003 dezembro 03.txt	AD
149	Competitividade	37	Impacto da tecnologia de informação na gestão de pequenas empresas	artigo 2000 jan 4.txt	AD
66	corporações transnacionais	17	Observatório da cidadania: monitorando as políticas públicas em âmbito global	artigo 1999 mai 5.txt	AD
931	Cultura Corporativa	285	Cultura Organizacional no Processo de Inteligência Competitiva	Artigo 2004 agosto 02.txt	AD
36	Cultura da organização	10	A inteligência competitiva e a área de informação tecnológica no Instituto de Pesquisas Tecnológicas do Estado de São Paulo S.A.	artigo 1999 jan 9.txt	AD
894	Cultura Organizacional	275	O Processo de Inteligência Competitiva em Organizações	Artigo 2003 junho 03.txt	AD
894	Cultura Organizacional	285	Cultura Organizacional no Processo de Inteligência Competitiva	Artigo 2004 agosto 02.txt	AD
460	Fatores críticos de sucesso	134	Motivações e fatores críticos de sucesso para o planejamento de sistemas interorganizacionais na sociedade da informação	artigo 2003 mai 8.txt	AD

O tratamento das palavras-chave foi realizado pelo gestor, com observação ativa do autor deste projeto. Por tratamento de palavras entende-se: o tratamento de gênero, a eliminação, quando possível, de sinônimos, de plural e de palavras-chave que não correspondem a termos propícios para a indexação e, em alguns casos, a substituição da palavra-chave por outra, de acordo com a cultura da área de Ciência da Informação e com o conhecimento do usuário. Este processo aconteceu de forma recursiva como um todo, ou seja, a partir do primeiro relatório as palavras foram sendo tratadas e novos relatórios foram emitidos com novas sessões de tratamento, até que se chegou a um ponto considerado satisfatório para o prosseguimento do projeto.

Observou-se, mais uma vez, que o gestor espalhava as folhas do relatório de um mesmo grupo de palavras e procurava ter uma visão do todo para seu tratamento. O título dos artigos auxiliou na decisão de impasses no tratamento.

A tabela de 'Palavras tratadas' ficou com um total de 730 palavras. Seguindo a carga desta tabela, foram criadas novas relações no banco de dados: 'Palavra Tratada' e o relacionamento 'Chave X Tratada'. A primeira refere-se às palavras resultantes do tratamento dado às palavras-chave e a segunda, ao par de equivalência entre palavra-chave e Palavra Tratada. Duas relações foram estabelecidas entre a tabela Artigo e a nova tabela Palavra Tratada. A primeira delas, 'Artigo Palavra Tratada', contém as freqüências de cada palavra tratada em cada artigo da base, através da somatória dessas freqüências. A segunda - 'Ocorrência Palavra Tratada' - foi apenas uma verificação para saber se as palavras tratadas encontrariam correspondência direta na base de artigos. Foi grata a surpresa: a relação 'Artigo Palavra Tratada' armazenou a freqüência de 648 palavras tratadas em pelo menos uma das unidades de texto

através da somatória da freqüência das palavras-chave a elas relacionadas. A verificação da freqüência da própria palavra tratada resultou no total de 643 palavras tratadas, o que denota a qualidade do tratamento realizado nas palavras-chave.

O modelo de entidades e relacionamentos do projeto passou a ser o especificado na Figura 11.

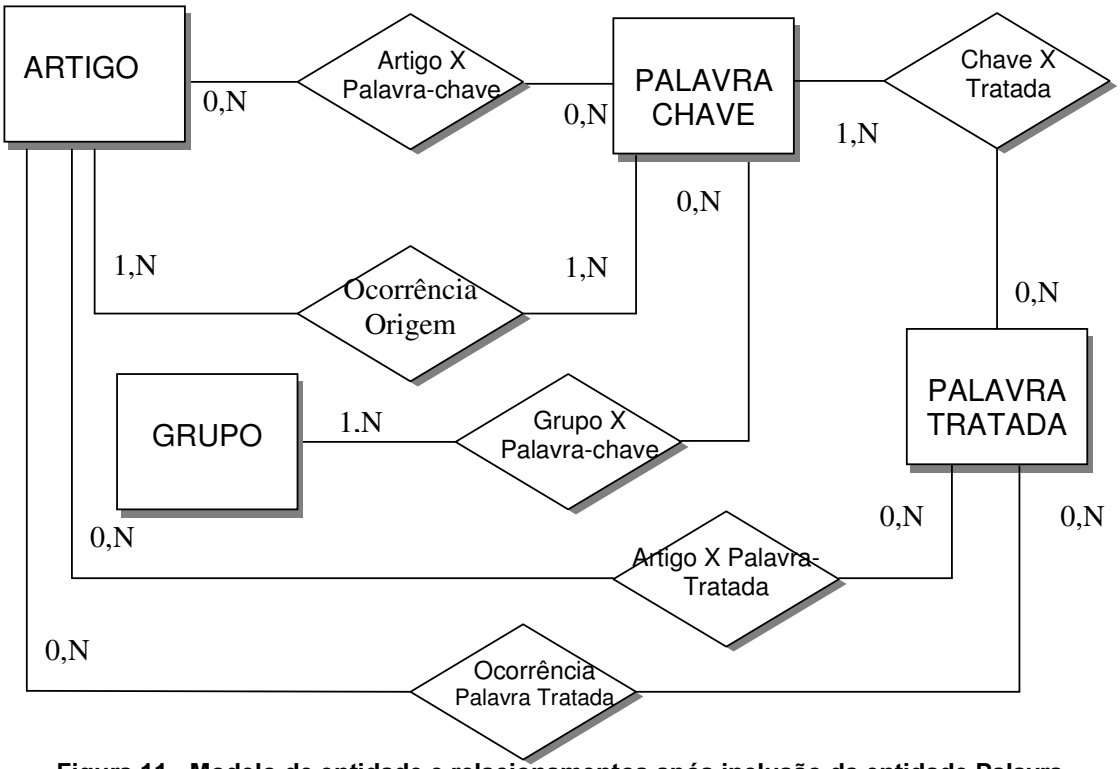

Figura 11 - Modelo de entidade e relacionamentos após inclusão da entidade Palavra Tratada

A equivalência entre Palavra Tratada e palavra-chave apontou que 573 palavras tratadas possuíam, cada qual, correspondência com apenas uma palavra-chave distinta. Durante a fase de tratamento foram excluídas 47 palavras-chave, relacionadas na Tabela 6.

Tabela 6 - Relação de palavras-chave excluídas

Código	Palavra-chave
12	Redes de informação corporativas
35	O profissional como agente de mudanças
41	Preservação da vontade sociocultural
57	Demanda da dimensão sociopolítica
58	Oferta técnico-científica
80	Monitoração em ciência e tecnologia/indicadores de competitividade
94	Estrutura de cursos para usuários de bases de dados
107	Serviços técnicos
181	Pensamento transversal
186	Sobrecarga informativa
218	Tríplice hélice
288	Diversidade humana
303	IMPA
307	Harvest
352	Produtividade de autores
373	Novas tecnologias
377	ANFIS
380	Cooperação e compartilhamento
448	Edgar Morin
449	Gaston Bachelard
472	Revista de divulgação científica
494	Personal Brain
530	Quissamã (Rio de Janeiro, Brasil)
549	Iniciativas internacionais
550	Propostas brasileiras
569	IDAMS
601	Henry Walter Bates
619	Impacto das Novas Tecnologias
624	Diversidade Disciplinar
625	Recomendações Curriculares
636	Ementa
652	UCC 2B
653	UCITA
662	Regulamentação das Novas Tecnologias
724	Visão de futuro
764	Estudos Sociais da Ciência
796	Gráficos Conceituais
806	Curso em informação
838	Obsessão Social
847	Recontextualização
849	Parâmetros de ordem
878	Análise funcional do comportamento verbal
896	Sociedade de Risco
898	Percepção de Risco
899	Compreensão Pública da Ciência e Tecnologia
932	Estratégias de Preservação
942	Explicação

3.3 Processamento estatístico

Foram realizados todos os cálculos estatísticos no ambiente e gerados os principais índices, seguindo as rotinas abaixo. Coube ao usuário gestor apenas a autorização do processo.

3.3.1 Cálculo das freqüências

A palavra freqüência, no âmbito da estatística, está associada ao número de vezes que determinado fato é observado (ANDERSON, et al., 2002; BUSSAB e MORETIN, 2002). Neste trabalho, considerou-se como freqüência ao número de vezes em que a palavra é encontrada nas unidades de texto.

Para armazenar a freqüência das palavras-chave foi introduzida a entidade 'Freqüência' no modelo de dados que passou a ter a estrutura apresentada na Figura 12. Optou-se por verificar a freqüência de cada palavra-chave em cada coleção da amostra, o que justifica uma palavra-chave ter mais de uma freqüência a ela relacionada.

Para armazenamento de dados relativos à freqüência das palavras tratadas, foram incluídas, na respectiva tabela, colunas específicas para este dado em cada unidade de texto, calculadas a partir da freqüência das palavras-chave correspondentes. Para este cálculo utilizou-se a linguagem *Structured Query Language* (SQL).

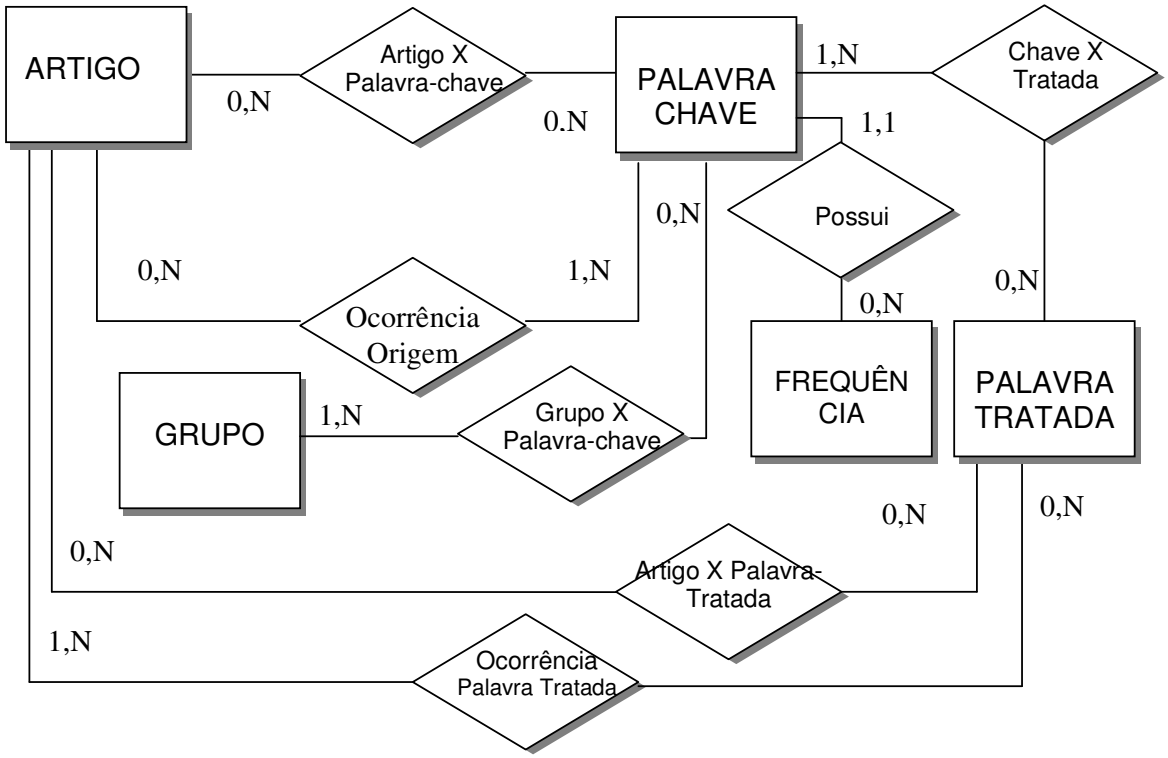

Figura 12 - Modelo de entidade e relacionamentos após inclusão da entidade Freqüência de Palavra Tratada

3.4 Verificação de termos no Tesauro Ideal

Para verificar a existência das palavras tratadas no TCI, foram criadas as entidades 'TCI TERMO' e 'TCI FACETA', contendo respectivamente os termos e as facetas existentes no TCI, bem como, os relacionamentos entre os termos neste tesauro - 'TCI Relacionamento'. A primeira recebeu os termos descritores e os não-descritores do TCI e a distinção entre eles foi feita por faixa de código. O modelo de entidades e relacionamentos do modelo passou a ter a configuração apresentada na Figura 13.

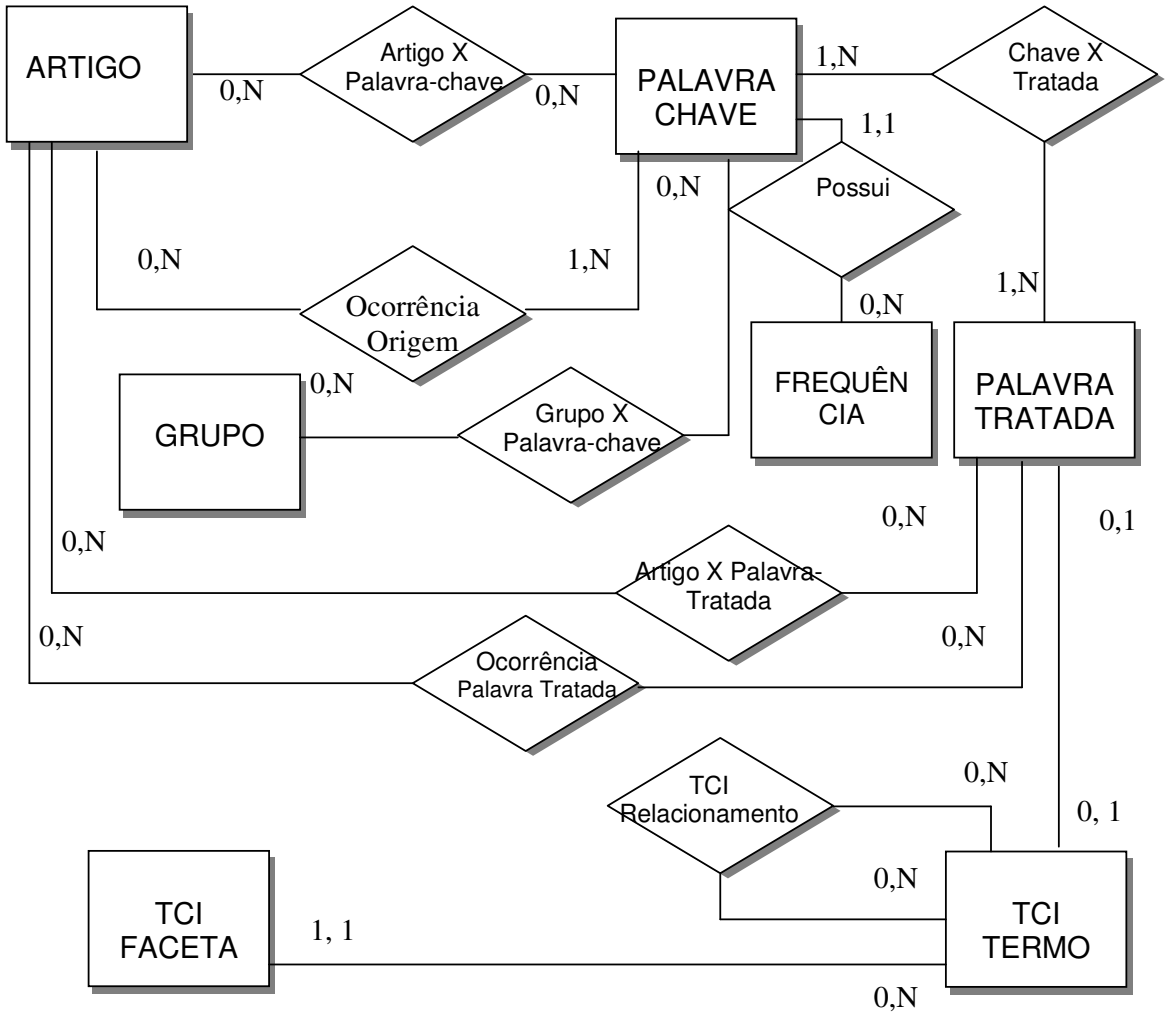

Figura 13 - Modelo de entidades e relacionamentos após inclusão das tabelas do TCI

Primeiramente, foram encontradas 161 ocorrências de palavras tratadas no TCI. Os dados foram analisados para verificar se existiam desencontros de singular-plural ou grafias diferentes para os mesmos termos. Para isso foi elaborada uma rotina que pesquisava coincidências de letras em partes específicas das palavras tratadas com as do TCI. O resultado mostrou 18 palavras tratadas nesta situação, listadas na Tabela 7.

Estas palavras foram consideradas como existentes no TCI, totalizando 179 ocorrências de palavras tratadas na tabela 'TCI Termo' correspondendo a 24,52%.

Tabela 7 - Palavras tratadas coincidentes em parte com as do TCI

Palavra tratada	Palavra no TCI
Administração de empresa	Administração de empresas
Análise documentária	Análise documental
Arquitetura de Sistema de Informação	Arquitetura de sistema
Base de dados bibliográfica	Base de dados bibliográficos
Biblioteconomia	Biblioteconomia (área de conhecimento)
Desenvolvimento de coleções	Desenvolvimento de coleção
Desenvolvimento de sistemas de informação	Desenvolvimento de sistema
Gestão de documento	Gestão de documentos
Informação institucional	Informação de empresas
Informação organizacional	Informação de empresas
Informação sociedade	Informação e sociedade
Interação homem-máquina	Interação homem máquina
Processamento de informação	Processamento da informação
Profissional da informação	Profissional de informação
Projeto de pesquisa e desenvolvimento	Projeto de pesquisa
Recuperação da informação	Recuperação de informação
Sistema de recuperação da informação	Sistema de recuperação de informação
Tecnologia da informação e comunicação	Tecnologia da informação e da comunicação

As etapas 'Seleção de termos' e 'Geração e atualização de tesauros' serão apresentadas no capítulo 4, a seguir.

4 Apresentação e análise dos dados

4.1 A base de artigos

Conforme descrito no capítulo 3, a amostra foi composta de 300 artigos, pertencentes a duas coleções. A Tabela 8 apresenta o total de textos classificados por coleção e ano de publicação, dados ilustrados pela Figura 14.

Tabela 8 - Quantidade de artigos por ano em cada coleção

Ano	Ciência da Informação	Datagrama Zero
1999	32	5
2000	26	16
2001	24	23
2002	31	24
2003	36	23
2004	32	28
Total	181	119

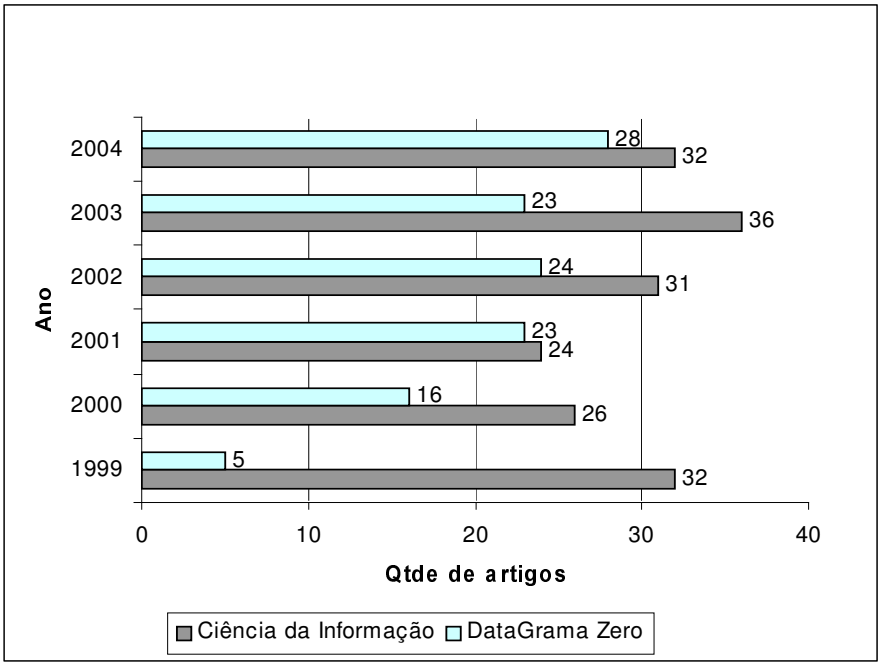

Figura 14 - Quantidade de artigos por ano em cada coleção

Percentualmente, 60,33% dos textos pertencem à revista *Ciência da Informação* e 39,67% à revista *Datagrama Zero*. Se classificada por ano, a freqüência de artigos tem um crescimento linear, sendo que os anos de 2002, 2003 e 2004 são aqueles em que há maior ocorrência de textos. É importante lembrar que foram colhidos apenas textos em língua portuguesa (Figura 15).

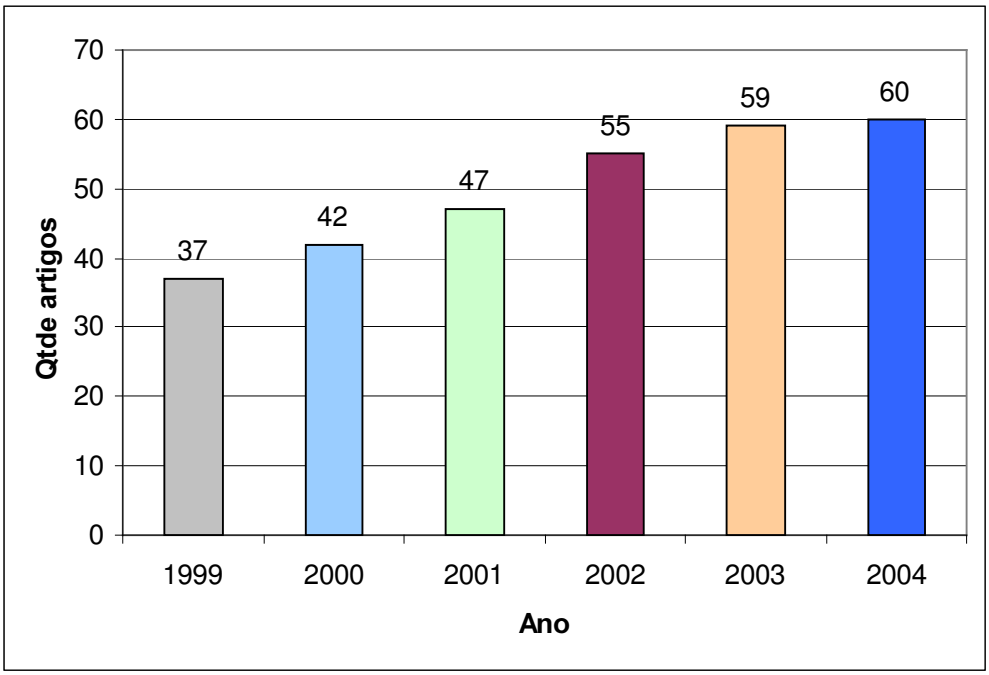

Figura 15 - Quantidade de artigos por ano

4.2 A base de palavras-chave

A base foi constituída por 957 palavras-chave. Destas, 523 ocorreram apenas em textos da revista *Ciência da Informação*; 349 na revista *Datagrama Zero* e 85 palavras ocorreram nas duas revistas, simultaneamente. A Figura 16 apresenta esses números em percentuais.

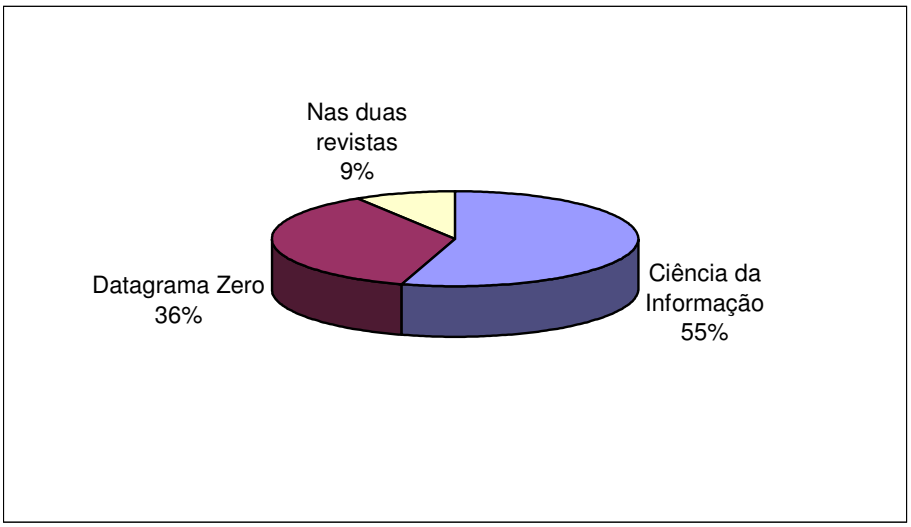

Figura 16 - Distribuição percentual da ocorrência de palavras-chave nas coleções

A ocorrência de palavras-chave em seu texto de origem está apresentada na Tabela 9, organizada por coleção. Ocorreram em mais de um texto 90 palavras-chave na revista *Ciência da Informação* e 59 na revista *Datagrama Zero*. Apenas 8 artigos, sendo quatro de cada coleção, não possuíam ocorrências de suas próprias palavras-chave em quaisquer das unidades de texto. Quando o conjunto de palavras-chave através da base formada foi verificado no conjunto de artigos, nenhum artigo apresentou-se nessa situação.

Também foi observado que 289 palavras-chave não ocorrem em quaisquer das unidades do texto, sendo que destas, 164 da revista Ciência da Informação e 125 da revista Datagrama Zero. Era previsível a não ocorrência de palavras-chave nas unidades do texto.

Tabela 9 - Quantidade de artigos com a quantidade de palavras-chave em cada coleção

Coleção	Quantidade de Artigos	Quantidade de Palavras
CI	2	1
	9	2
	43	3
	41	4
	47	5
	20	6
	8	7
	6	8
	2	9
	1	10
	1	11
	1	19
Total	**181 artigos**	
DATAGRAMA	5	2
	19	3
	21	4
	48	5
	15	6
	8	7
	1	8
	1	9
	1	10
Total	**119 artigos**	

Observou-se que, em alguns casos e para ambas as revistas, as regras de submissão no tocante ao número de palavras-chave não haviam sido respeitadas, embora a maior concentração de palavras-chave por artigo tenha ocorrido na faixa de três a seis palavras.

4.3 Freqüência de palavras-chave nas unidades de seus textos de origem

Foi verificada a ocorrência das palavras-chave em cada unidade de seu texto de origem. Esses dados são apresentados na Tabela 10.

Tabela 10 - Artigos com ocorrência de suas próprias palavras-chave nas unidades de texto

	Total de artigos	no Título	no Resumo	no Texto	na Bibliografia
Datagrama Zero	119	95 (79,84%)	104 (87,39%)	113 (94, 96%)	72 (60,50%)
Ciência da Informação	181	140 (77,35%)	161 (88,95%)	176 (97,28%)	126 (69,62%)
Total	300	235 (78,34%)	265 (88,34%)	289 (96,34%)	198 (66,00%)

Tomando como base a revista *Datagrama Zero*, utilizada no pré-teste, nota-se uma melhoria dessa freqüência. Observa-se que novos textos, não disponíveis no instante do pré-teste, foram inseridos na amostra. Para comparar com o pré-teste, a Figura 17 apresenta a freqüência de palavras-chave no título de seu artigo de origem para as palavras da revista *Datagrama Zero*.

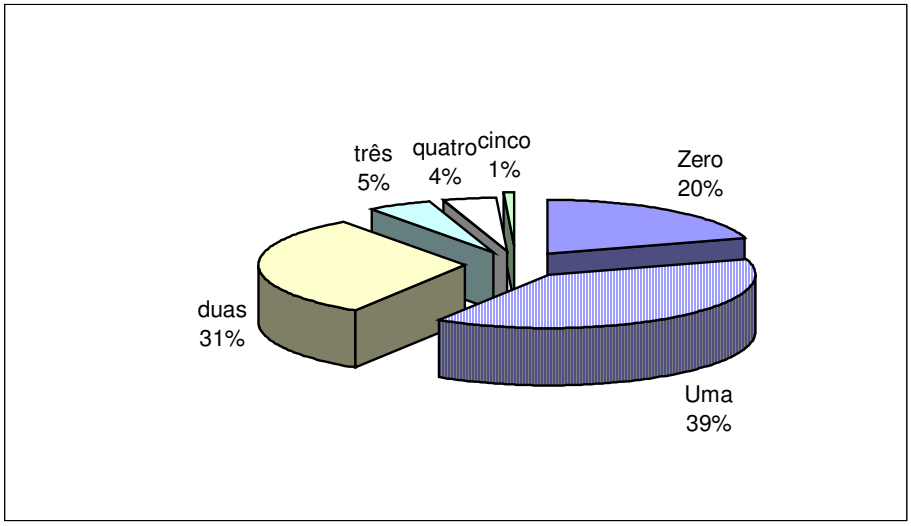

Figura 17 - Percentual de textos da coleção Datagrama Zero com a respectiva quantidade de palavras-chave em seu título

Na Figura 18, apresenta-se a distribuição para a coleção *Ciência da Informação*, percebida como mais uniforme.

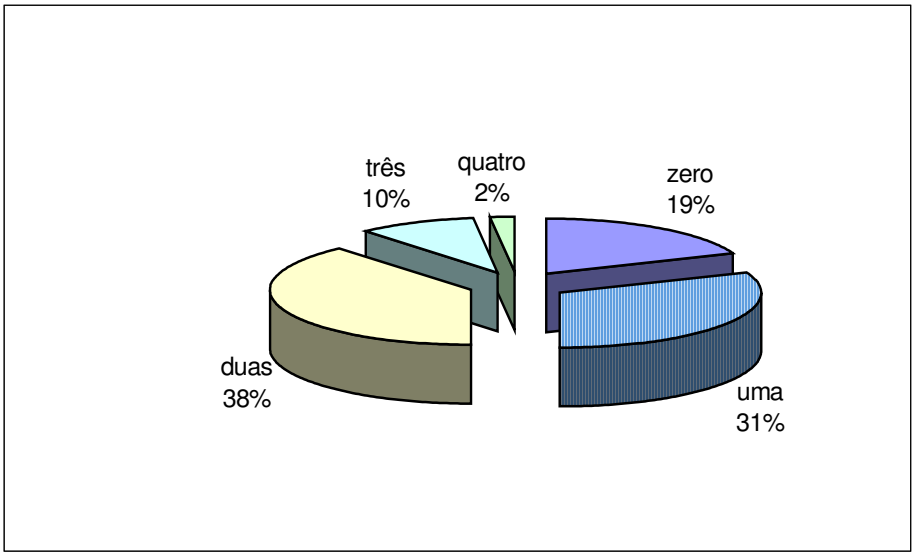

Figura 18 - Percentual de textos da coleção Ciência da Informação com a respectiva quantidade de palavras-chave em seu título

Para cada palavra-chave, foi observada a freqüência em cada unidade de texto. Como são unidades distintas, as tabelas 11, 12, 13 e 14 apresentam respectivamente, a freqüência total de palavras-chave nos títulos, nos resumos, nos textos e nas bibliografias de seu artigo de origem. Para que seja possível uma comparação entre estes valores, as freqüências serão também apresentadas em percentuais.

Tabela 11 - Quantidade de palavras-chave por faixa de freqüência no título do texto de origem

Quantidade de palavras-chave	Faixa de freqüência	Percentual
680	zero	71,06%
273	de 1 a 5	28,53%
1	de 6 a 10	0,10%
1	de 11 a 15	0,10%
2	de 16 a 20	0,21%

Tabela 12 - Quantidade de palavras-chave por faixa de freqüência no resumo do texto de origem

Quantidade de palavras-chave	Faixa de freqüência	Percentual
518	Zero	54,13%
406	de 1 a 5	42,42%
24	de 6 a 10	2,51%
3	de 11 a 15	0,31%
4	de 16 a 20	0,42%
2	acima de 20	0,21%

Tabela 13 - Quantidade de palavras-chave por faixa de freqüência no corpo do texto de origem

Quantidade de palavras-chave	Faixa de freqüência	Percentual
285	Zero	29,78%
284	de 1 a 5	29,68%
132	de 6 a 10	13,79%
59	de 11 a 15	6,17%
44	de 16 a 20	4,60%
153	Acima de 20	15,98%

Tabela 14 - Quantidade de palavras-chave por faixa de freqüência na bibliografia do texto de origem

Quantidade de palavras-chave	Faixa de freqüência	Percentual
717	Zero	74,92%
191	De 1 a 5	19,96%
31	De 6 a 10	3,25%
5	De 11 a 15	0,52%
5	De 16 a 20	0,52%
8	Acima de 20	0,83%

4.4 Freqüência de palavras-chave da base nas unidades de texto

A Tabela 15 apresenta a freqüência das palavras-chave quando os artigos são averiguados com toda a base de palavras-chave formada. Os valores anteriores aproximam-se de 100%, confirmando e melhorando o que já havia sido visto no pré-teste. Esta melhoria deve-se ao fato de terem sido encontradas as palavras-chave em ambas as coleções.

Tabela 15 - Ocorrências nos artigos das palavras-chave da base formada

Freqüência de artigos com ocorrências das palavras-chave da base formada					
	Total de artigos	No Título	No Resumo	No Texto	Na Bibliografia
Datagrama Zero	119 (100%)	116 (97,48%)	119 (100%)	119 (100%)	118 (99,15%)
Ciência da Informação	181 (100%)	179 (98,90%)	181 (100%)	181 (100%)	181 (100%)
Total	300	295 (98,34%)	300 (100%)	300 (100%)	299 (99,67)

As tabelas 16 a 19 apresentam, respectivamente, a freqüência total dessas palavras nos títulos, nos resumos, nos textos e nas bibliografias dos artigos.

Tabela 16 - Palavras-chave nos títulos dos textos por faixa de freqüência

Quantidade de palavras-chave	Faixa de freqüência	Percentual
615	Zero	64,26%
308	de 1 a 5	32,18%
20	de 6 a 10	2,09%
4	de 11 a 15	0,42%
3	de 16 a 20	0,31%
7	acima de 20	0,73%

Tabela 17 - Palavras-chave nos resumos dos textos por faixa de freqüência

Quantidade de palavras-chave	Faixa de freqüência	Percentual
419	Zero	43,78%
407	de 1 a 5	42,54%
44	de 6 a 10	4,60%
22	de 11 a 15	2,30%
18	de 16 a 20	1,88%
47	Acima de 20	4,90%

Tabela 18 - Palavras-chave no corpo dos textos por faixa de freqüência

Quantidade de palavras-chave	Faixa de freqüência	Percentual
172	zero	17,97%
213	de 1 a 5	22,26%
104	de 6 a 10	10,87%
69	De 11 a 15	7,21%
42	De 16 a 20	4,39%
357	Acima de 20	37,31%

Tabela 19 - Palavras-chave nas bibliografias dos textos por faixa de freqüência

Quantidade de palavras-chave	Faixa de freqüência	Percentual
494	zero	51,62%
214	de 1 a 5	22,36%
60	de 6 a 10	6,27%
29	de 11 a 15	3,03%
17	de 16 a 20	1,78%
143	Acima de 20	14,94%

Os dados confirmaram o pré-teste. A reunião de palavras-chave de coleções de artigos em uma única base pode aumentar o poder de recuperação dos documentos. É preciso ressaltar que as palavras-chave de uma coleção foram encontradas também nos artigos da outra coleção, reafirmando o que se esperava em relação ao reunir textos científicos de uma mesma área do conhecimento.

4.5 Trabalhando com palavras tratadas

O total de linhas na tabela de palavras tratadas foi de 730. A relação ‘Chave X Tratada’ garantiu que uma palavra tratada ficasse associada a pelo menos uma palavra-chave e vice-versa. A relação ‘Artigo X Palavra Tratada’ recebeu dados através da navegação entre as relações ‘Artigo X Palavra-Chave’ e ‘Chave X

Tratada'. Desta forma, foram armazenadas todas as freqüências das palavras tratadas para cada unidade de texto, em todos os artigos.

É necessário lembrar que, embora o total de palavras tratadas seja de 730, estão relacionadas às 957 palavras-chave, das quais 289 não haviam sido observadas nas unidades de textos. Este fato fez com que 82 palavras tratadas não constassem em 'Artigo X Palavra Tratada'. Porém, em todos os artigos foram encontradas as palavras tratadas e o menor número foi de trinta palavras tratadas por artigo, sendo o maior de 127.

Foi realizado um experimento à parte para averiguar se a própria palavra tratada ocorria nas unidades de texto da base de artigos. Seu resultado foi armazenado na relação 'Ocorrência Palavra Tratada'. O resultado superou a freqüência das palavras-chave. Foram encontradas palavras tratadas em todos os artigos, sendo que o menor número de palavras por artigo foi de 32 e o maior de 132 palavras. Porém, enquanto na relação 'Artigo X Palavra Tratada' não houve freqüência para 82 palavras tratadas, na relação 'Ocorrência Tratada' a não-ocorrência foi de 87 palavras tratadas.

Observa-se que a relação 'Ocorrência Palavra Tratada' foi criada apenas como termômetro do tratamento dado às palavras-chave e não será utilizada nas estatísticas deste trabalho.

4.6 Cálculo do desvio padrão e do escore padronizado das freqüências

Neste trabalho, as freqüências das palavras tratadas foram obtidas em unidades distintas do texto: título, resumo, corpo do texto e bibliografia. Desejava-se obter a freqüência total da palavra tratada no artigo, mas esta

freqüência não poderia vir da simples somatória das freqüências nas unidades do texto, pois cada uma delas tem sua estrutura de formação e suas regras. Tomando como grão a palavra, observa-se que as unidades de texto, por sua própria natureza, possuem ocorrências de palavras distintas. Os textos científicos possuem regras próprias e mesmo que elas sejam burladas, jamais se construirá um resumo com o mesmo número de palavras de um texto. A comparação fica ainda mais sem sentido se tomarmos o título e compara-lo com qualquer uma das outras unidades do texto.

Ressalta-se também que as unidades de texto possuem valores subjacentes a ela agregados, não abordados neste trabalho. Tomando em conta estes valores, as palavras constantes em um título teriam um peso maior que as encontradas no corpo do texto, por exemplo.

Optou-se pela solução do problema apoiado na teoria estatística, onde para se comparar unidades de medidas de diferentes conjuntos de dados, utiliza-se o escore padronizado. O resumo deste cálculo é apresentado na Tabela 20 e o Anexo 2 traz maiores detalhes do mesmo.

Tabela 20 - Resumo de cálculo média, mediana, desvio padrão dos escores padronizados

Análise Descritiva do Índice	
n	730
Média	0,00000
Mediana	-0,26001
Desvio Padrão	0,90204
Mínimo	-0,31523
Máximo	11,58814

Para facilitar a apresentação desses índices em histograma, foram criados dez percentis e verificada a freqüência dos índices por faixa de valores. As Tabelas 21 e 22 apresentam, respectivamente, os percentis criados e sua freqüência em faixas de valores, com os quais foi possível traçar o histograma da Figura 19.

Tabela 21 - Percentis de valores de índices calculados

Percentis	
10	-0,31549
20	-0,30667
30	-0,29535
40	-0,28032
50	-0,26007
60	-0,23471
70	-0,17329
80	-0,03399
90	0,56607
100	11,61334

Tabela 22 - Freqüência na faixa de valores de índices

Faixa de valores	Freqüência
Igual a –1	0
de –1 a 0	594
de 0 a 1	90
de 1 a 2	27
de 2 a 3	7
de 3 a 4	6
de 4 a 5	2
de 5 a 6	2
de 6 a 7	0
de 7 a 8	0
de 8 a 9	0
de 9 a 10	0
de 10 a 11	0
de 11 a 12	2

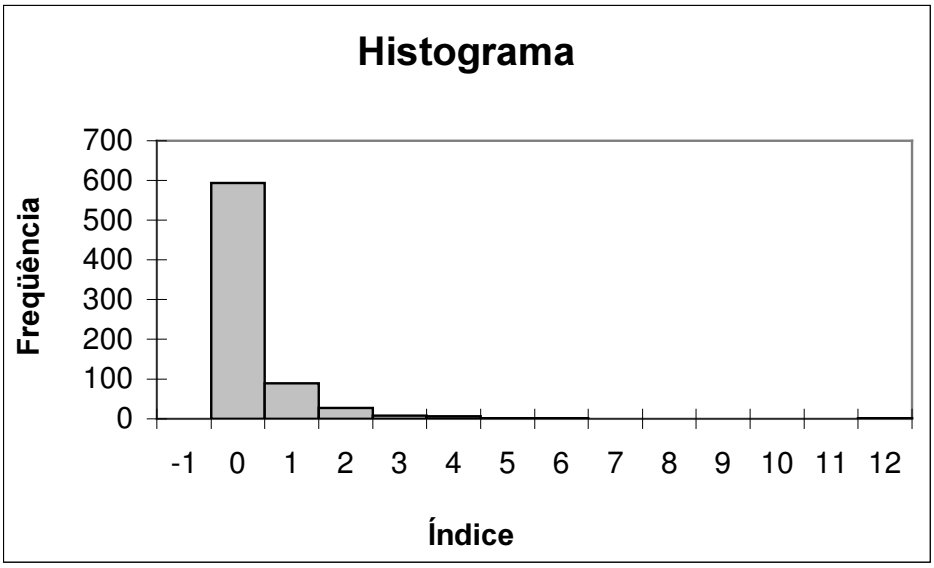

Figura 19 - Histograma das faixas de valores de índices

Estes índices foram divididos em três faixas de acordo com seu valor: ‘menor que zero’, ‘maior ou igual a zero e menor que 1’ e ‘maior ou igual a 1’. Como foram calculados a partir da freqüência das palavras tratadas, são significativos, respectivamente, de palavras que podem indicar novidade na área ou palavras que voltaram a ser utilizadas nos artigos, e por isso uma freqüência mais baixa no período; de palavras existentes em todo o período e com uma distribuição mais homogênea; e de palavras muito utilizadas e portanto pertencentes a temas muito discutidos no período, ou palavras consolidadas na área e por isso com alta freqüência. A Figura 20 mostra os percentuais de distribuição destes índices nestas faixas.

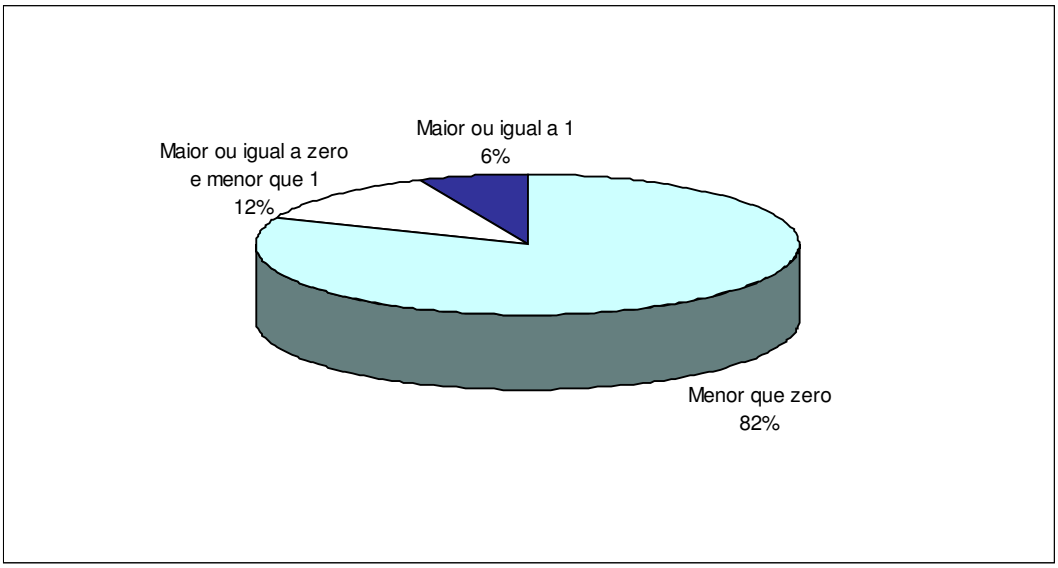

Figura 20 - Percentual do índice de palavras tratadas

4.7 Palavras tratadas não existentes no TCI

O objetivo do ambiente era indicar quais palavras-chave deveriam fazer parte do TCI. Para isto, a base de termos do TCI foi transportada para uma tabela no gerenciador de bancos de dados MySql. Verificou-se que 179 palavras tratadas já existiam no TCI, correspondendo a 24,52% do total, mas interessavam mais as 551 (75,48%) restantes, número que, comparado a outros já tratados neste trabalho, revela alguns pontos, discutidos adiante. A relação completa das palavras tratadas, com indicativo das que ocorreram e das que não ocorreram no TCI é apresentada no Anexo 3.

A Tabela 23 traz o percentual, por ano, das palavras-chave não existentes no TCI, dado obtido pelo relacionamento existente entre palavras-chave e palavras tratadas - 'Chave Tratada' - e o conjunto formado de palavras tratadas não existentes no TCI.

Tabela 23 - Percentual de palavras-chave não existentes no TCI por ano do artigo de origem

Ano	Quantidade de Palavras-chave não existentes no TCI	Percentual
1999	110	13%
2000	130	15%
2001	130	15%
2002	139	16%
2003	187	22%
2004	158	19%
TOTAL	854	100%

Os artigos da base deste trabalho contemplam apenas seis anos de produção na Ciência da Informação e por apenas duas revistas, devido à limitação da disponibilidade em meio digital e ao acesso público. A produção do passado e o percurso desta ciência estão retratados no TCI, já que foram considerados nos tesauros que lhe serviram de fonte; além da legitimidade através dos órgãos que os formaram, dos objetivos para os quais foram criados e da utilização de cada um deles por seu nicho de usuários. Assim, de alguma forma comparou-se um tesauro preparado para artigos científicos produzidos antes do final dos anos 90 com artigos científicos produzidos posteriormente. Era esperado um número baixo de palavras tratadas existentes no tesauro. O fato de se ter encontrado um número próximo a 25% delas no TCI pode revelar alguns fatos:

- a produção em Ciência da Informação oscila pela própria interdisciplinaridade da área, fazendo com que temas envolvendo mais de uma área do conhecimento sejam discutidos à luz de suas bases teóricas e tenham produção científica privilegiada conforme apareçam em destaque, na aproximação verificada. A Tabela 24 apresenta as palavras-chave não

existentes no TCI classificadas nos grupos que as tornaram possíveis. O percentual de palavras do grupo 'Limítrofe' revela a interdisciplinaridade da área. A Tabela 25 detalha por grupo / ano este percentual;

Tabela 24 - Palavras-chave não existentes no TCI por grupo de classificação para tratamento

Grupo	Quantidade de Palavras-chave não existentes no TCI	Percentual
Ciências Médicas	1	0,12%
História	1	0,12%
Psicologia	1	0,12%
Ciências Humanas	2	0,23%
Ciências Sociais	2	0,23%
Filosofia	3	0,35%
Museologia	3	0,35%
Arquivologia	6	0,70%
Linguística	11	1,29%
Ciência Cognitiva	23	2,69%
Ciências Políticas	24	2,81%
Sociologia	24	2,81%
Educação	27	3,16%
Ciência da Computação	39	4,57%
Administração	69	8,08%
Biblioteconomia	78	9,13%
Limítrofe	213	24,94%
Ciência da Informação	327	38,29%
Totais	854	100,00%

Tabela 25 - Percentual de Palavras-chave não existentes no TCI por Grupo / Ano

Quantidade de Palavras-chave não existentes no TCI por Grupo							
Grupo	Ano	Quantidade Palavras-chave	%	Grupo	Ano	Quantidade Palavras-chave	%
Administração	1999	17	1,99%	Ciências Políticas	1999	5	0,59%
	2000	5	0,59%		2000	6	0,70%
	2001	15	1,76%		2001	3	0,35%
	2002	5	0,59%		2002	1	0,12%
	2003	19	2,22%		2003	4	0,47%
	2004	8	0,94%		2004	5	0,59%
Arquivologia	1999	3	0,35%	Ciências Sociais	2001	1	0,12%
	2003	1	0,12%		2004	1	0,12%
	2004	2	0,23%	Educação	1999	1	0,12%
Biblioteconomia	1999	13	1,52%		2000	5	0,59%
	2000	7	0,82%		2001	1	0,12%
	2001	14	1,64%		2002	13	1,52%
	2002	16	1,87%		2003	6	0,70%
	2003	18	2,11%		2004	1	0,12%
	2004	10	1,17%	Filosofia	2000	1	0,12%
Ciência Cognitiva	1999	5	0,59%		2002	1	0,12%
	2000	3	0,35%		2004	1	0,12%
	2001	4	0,47%	História	2000	1	0,12%
	2002	2	0,23%	Limítrofe	1999	25	2,93%
	2003	4	0,47%		2000	33	3,86%
	2004	5	0,59%		2001	29	3,40%
Ciência da Computação	1999	1	0,12%		2002	31	3,63%
	2000	7	0,82%		2003	47	5,50%
	2001	8	0,94%		2004	48	5,62%
	2002	6	0,70%	Linguística	1999	2	0,23%
	2003	12	1,41%		2000	1	0,12%
	2004	5	0,59%		2001	3	0,35%
Ciência da Informação	1999	38	4,45%		2002	2	0,23%
	2000	54	6,32%		2003	2	0,23%
	2001	46	5,39%		2004	1	0,12%
	2002	59	6,91%	Museologia	2004	3	0,35%
	2003	66	7,73%	Psicologia	2003	1	0,12%
	2004	64	7,49%	Sociologia	2000	6	0,70%
Ciências Humanas	2000	1	0,12%		2001	5	0,59%
	2003	1	0,12%		2002	3	0,35%
Ciências Médicas	2001	1	0,12%		2003	6	0,70%
					2004	4	0,47%
Totais		545	63,82%			309	36,18%

A Figura 21 apresenta o índice de freqüência equivalente ao escore padronizado para as palavras tratadas não existentes no TCI.

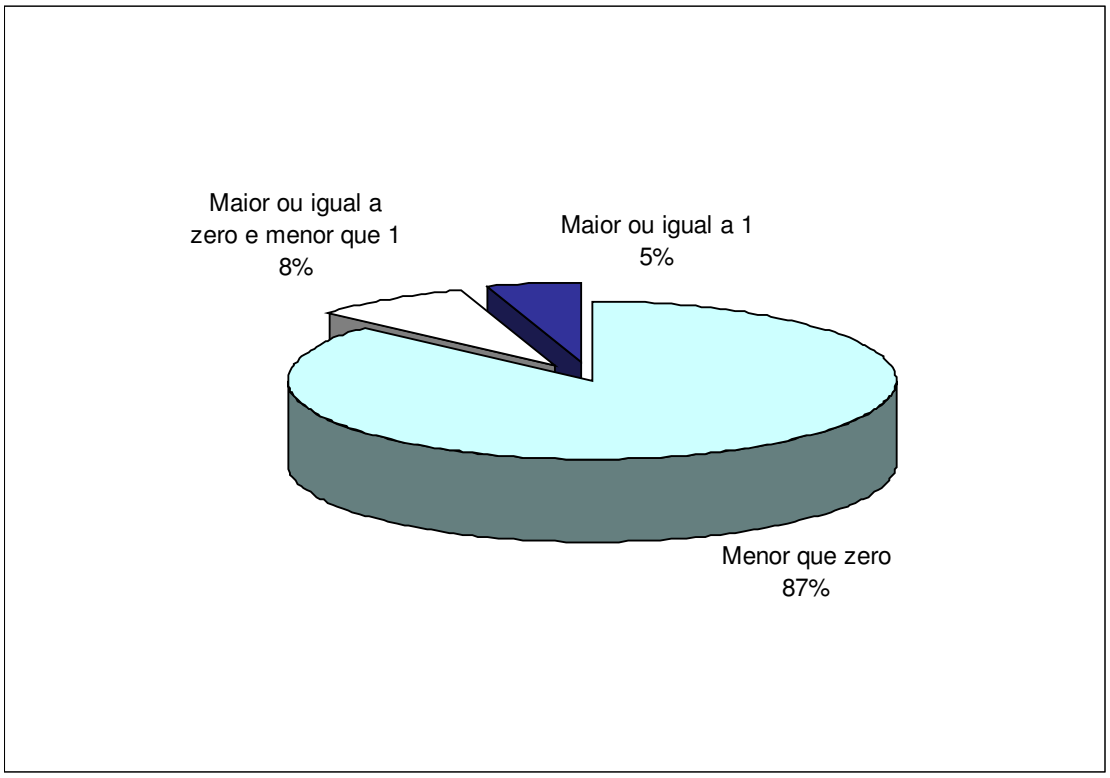

Figura 21 - Percentual do índice de palavras tratadas não existentes no TCI

- a Ciência da Informação parece estar sendo revisitada. Alguns temas da própria área são aqueles já estudados no passado. Mas, sua interdisciplinaridade com outras áreas e as contribuições advindas trazem novas oportunidades. Este trabalho é um exemplo disto e o percentual alto para o grupo 'Ciência da Informação' também revela isto (Tabela 24);
- nessa linha, o percentual tímido referente à lingüística aponta a retomada das discussões sobre terminologia, linguagens de recuperação de informação, sistemas de indexação, automação de resumos e de traduções, entre outros;

- o percentual não muito alto para a Ciência da Computação deve-se ao fato de que, durante o tratamento, as palavras-chave relacionadas à área de Tecnologia da Informação foram classificadas no grupo 'Limítrofe'. Mesmo assim, corresponderam a 4,58% do total e revelam principalmente textos que abordam a Web e sua adoção pela Ciência da Informação;
- o percentual de 2,7% para a Ciência Cognitiva juntamente com a curva formada pelos percentuais ano a ano mostra a interdisciplinaridade com o tema Conhecimento e as pesquisas que retornam sobre o tema na Ciência da Informação;
- a Arquivologia e a Museologia possuem um percentual muito pequeno. A interdisciplinaridade é notória nas áreas, mas a divulgação de trabalhos envolvendo Ciência da Informação e estes campos ainda se dá de forma separada, até mesmo como reforço à divisão do exercício profissional, tão polêmica em nossos dias;
- no final dos anos 90, muito se discutiu em torno do tema gerência da informação. O reflexo desta fase está nítido no ano de 1999 e na retomada do tema em textos contendo palavras-chave como gestão da informação, gestão da qualidade, inteligência competitiva, entre outros;
- o TCI apresentou necessidade de atualização em pontos estratégicos de outras áreas, para corresponder à realidade: um bom índice para a Ciências Políticas, que estiveram presentes em textos enfocando governo eletrônico, governabilidade, gestão participativa, informação e órgãos públicos, etc.; a Sociologia, com a chegada da discussão da inclusão digital; o alto índice para o grupo Biblioteconomia que teve durante o período diversos textos

tratando da aplicação de padrões de metadados e o início da discussão de bibliotecas digitais; e a Educação, com textos pertinentes à ensino e tecnologia, educação à distância e educação tecnológica.

A distribuição mantém-se para este grupo de palavras, porém um percentual menor na faixa 'maior ou igual a zero e menor que 1' reforça o fato destas palavras tratarem assuntos com freqüência mais homogênea na base de artigos. Assim como, a manutenção relativa do percentual de índices altos para palavras que denotam novidade ou assuntos que estão sendo revisitados.

4.8 Distância entre palavras tratadas

Foi verificada a distância entre duas palavras-chave que ocorriam, simultaneamente, em mais de um texto da base de artigo.

Um texto é composto de frases que juntas formam parágrafos. O texto científico possui subitens distintos, cada um deles formado de diversos parágrafos. Neste trabalho, considerou-se a distância entre palavras à menor distância de palavras distintas, nas seguintes métricas de texto: 'mesma frase', 'mesmo parágrafo', 'parágrafos diferentes' e 'itens diferentes'. São simétricas, ou seja, a distância entre a palavra *A* e a palavra *B* é igual à distância entre a palavra *B* e a palavra *A*, em qualquer circunstância.

Foi criada uma rotina específica para esta medição. Primeiramente, foi elaborada uma visão[23] de dados formada de pares de palavras que ocorriam em um mesmo texto mais de uma vez e cujas palavras-chave se relacionavam com as palavras tratadas não existentes no TCI, o que foi lido pelo programa.

[23] A palavra visão está sendo usada neste ponto como tradução do objeto *view* do modelo relacional

Para cada artigo da base onde ocorria o par de palavras-chave foram armazenados os dados sobre as distâncias na menor medida das métricas citadas anteriormente. Os resultados foram armazenados em uma tabela específica e seu resultado chegou a quase 187.000 linhas, conjugando um total de 107 palavras-chave que ocorriam em 1450 pares distintos.

Foram encontradas 45 palavras tratadas entre as não existentes no TCI correspondentes a estas 107. A Figura 22 mostra o percentual do índice de freqüência para estas palavras, onde observa-se que o maior índice está entre palavras consolidadas na área e que novas palavras, ou palavras que passam a ser novamente utilizadas, possuem baixa freqüência, justamente pelo tema se encontrar em fase de crescimento.

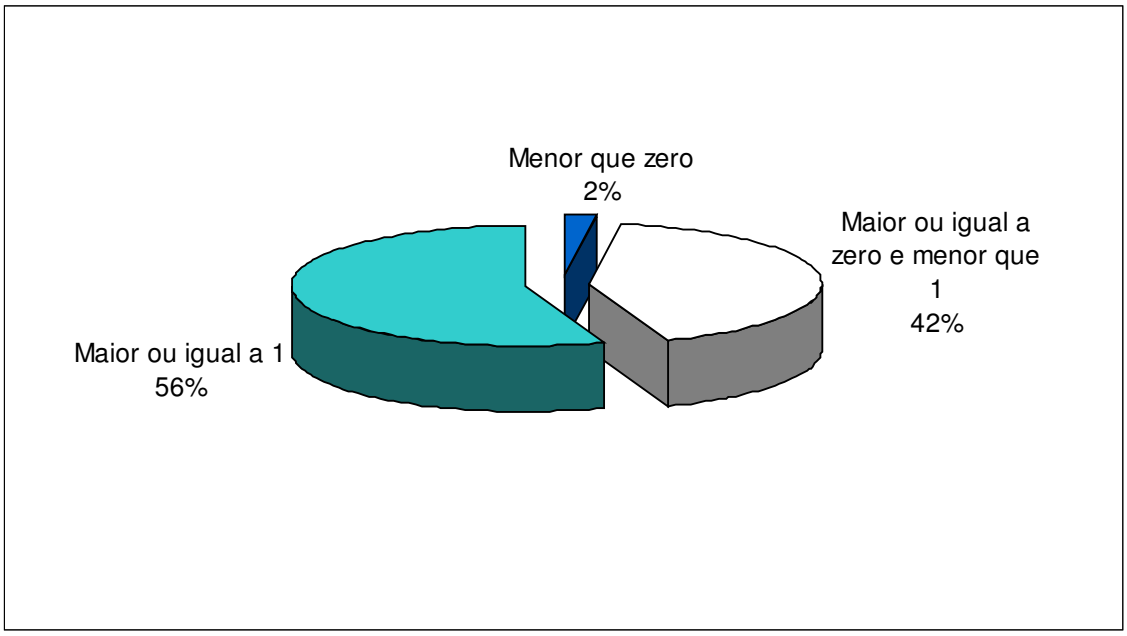

Figura 22 - Percentual de faixa de índice de palavras tratadas para palavras-chave que ocorrem em pares

4.7 Escolha dos termos

O ambiente previa a escolha dos termos para a atualização do tesauro. Todas as 551 palavras tratadas não encontradas no TCI precisam ser analisadas para esta inclusão. Um aplicativo foi montado para uma interação do usuário. Nele, estas palavras foram divididas conforme os grupos de valor de índice abordado anteriormente: 'menor que zero', 'maior ou igual a zero e menor que 1' e 'maior ou igual a 1', que passaram a receber o nome, respectivamente, de 'termos que denotam novidade ou retomada de termos'; 'termos contemporâneos' e 'termos estáveis'.

Observa-se que apesar de todo esforço computacional e estatístico deste trabalho, ele foi antes de tudo orientado pela preocupação de garantir o carater amigável da linguagem, levando em conta a terminologia, os aspectos semânticos e o contexto da atualização. Este princípio levou a indicação dos termos aos usuários gestores, sendo que a escolha da ordem de atualização e da inclusão ou não do termo no tesauro continuam sob sua responsabilidade.

O ambiente proporciona ao usuário, após a escolha de uma palavra tratada em um destes grupos, um quantitativo de artigos com a ocorrência desta palavra para cada ano do período dos artigos (1999-2004). São também mostrados os títulos de cada artigo onde ela aparece com a facilidade de acesso ao próprio artigo. O Anexo 4 apresenta exemplos destas telas.

5 Considerações finais

Este trabalho partiu da hipótese de que palavras-chave encontradas em textos científicos poderiam servir de base para geração e atualização de tesauros, tendo como suporte teórico as garantias literária, de uso e estrutural. Somando a elas, foi proposta a garantia da estrutura de texto.

A disseminação de textos científicos por meio eletrônico trouxe dinamismo à comunicação científica, mas evidenciou também a obsolescência das linguagens e das metodologias de indexação. A velocidade como novos textos são produzidos e disseminados fez com que o problema da atualização viesse à tona.

Como proposta de solução montou-se um ambiente para reunir e tratar as palavras-chave de textos científicos, coletadas e armazenadas em uma única base. Verificou-se que os textos científicos de uma mesma área possuem em suas unidades de texto palavras-chave de outros textos, sejam eles de sua própria coleção ou de outra coleção da mesma área.

O tratamento de palavras-chave proposto mostrou sua facilidade frente ao trabalho de identificação, de extração, de leitura e de escolha de termos comuns aos trabalhos de geração e atualização de tesauros. Contudo será necessário o cotejamento da lista resultante com as fontes específicas do domínio.

Foi construído um tesauro em Ciência da Informação, a partir da síntese de tesauros existentes, que estará disponível na Web para futuros trabalhos e

atualizações. Ao final do experimento observou-se que este carecia de termos que demonstram tendências de estudo no campo.

A tecnologia da computação e o cálculo estatístico foram empregados como forma de auxiliar o profissional de indexação na tarefa de manter atualizadas as ferramentas de indexação indicando e classificando termos inexistentes no tesauro em relação ao grau de novidade. O uso de procedimentos estatísticos e computacionais auxiliaram na tomada de decisão quanto ao caráter representativo dos termos.

As palavras-chave demonstraram ser um excelente indicativo das garantias literárias e de uso, sobretudo porque elas representam o substrato teórico indicado pelos próprios autores. Nesse caso, a redução do intervalo existente entre o processo de produção de conhecimento e a criação de instrumentos de representação e recuperação da informação fica facilitada a partir da extração e comparação automática da freqüência das palavras–chave num dado campo de conhecimento. Verificou-se que a atualização da linguagem tem, nessas palavras, um balizamento importante.

A ausência de adoção de padrões para a composição das palavras-chave revela que os autores comunicam-se entre si numa linguagem altamente especializada. Nesse sentido, é preciso eleger outros instrumentos mais gerais para a composição da base terminológica (dicionários especializados, sistemas de classificação, taxonomias, dentre outros) para que a linguagem gerada possa efetivar-se como um mapa conceitual representativo da área. É preciso que esse instrumento vá além das particularidades presentes no discurso de

um ou outro pesquisador e das marcas locais e temporais que, por vezes, impregnam o discurso científico.

O conjunto final de palavras tratadas demonstrou que a produção do conhecimento na área tem incidido sobre experiências interdisciplinares que articulam a Ciência da Informação com outros campos do conhecimento, o que leva sua representação a ser influenciada por essas experiências. Compreender como se dá o estabelecimento das interfaces entre a Ciência da Informação e os demais campos do conhecimento parece ser fundamental para viabilizar automaticamente a geração de tesauros ou produtos que automatizem parte desse processo, como o aqui apresentado.

A interdisciplinaridade entre as áreas de Ciência da Informação e a Ciência da Computação permitiu que métodos de ambas as áreas fossem empregados na solução do problema, apontando maiores possibilidades para futuros trabalhos nos quais o conhecimento específico de cada uma delas seja compartilhado em vistas ao objetivo. Verificou-se no projeto o compartilhamento das metodologias das áreas, possibilitando que o método computacional simulasse, de forma metafórica, os processos de indexação.

A disponibilidade de textos científicos em meio eletrônico crescerá a cada dia. O trabalho demarcou a importância da garantia da manutenção de estruturas pré-estabelecidas para os mesmos, de acordo com regras vigentes, correspondendo a facilidades do trabalho computacional para reconhecimento e extração de palavras.

O término do trabalho suscita continuidade. Recomenda-se que trabalhos futuros nessa mesma linha estejam principalmente atentos:

- à riqueza das referências bibliográficas encontradas nos textos científicos e ao que as palavras nela dispostas poderão revelar. O trabalho da bibliometria relacionado ao de geração e manutenção de tesauros pode revelar novidades para automação dos processos de indexação. Este trabalho abordou parte disso, mas há muito a ser feito, principalmente se relacionado à identificação de autores consagrados em sub-áreas específicas do conhecimento, ao relacionamento entre palavras de artigos e as existentes nos tesauros das áreas como subsídio aos futuros relacionamentos dos termos, entre muitos outros;

- aos trabalhos que partam para coleta de palavras em outras unidades de texto, como nos títulos dos itens de artigos científicos, relacionando estes às freqüências de palavras-chave no próprio item;

- ao ato de conjugar freqüência de palavras com seu significado semântico, quer quando vistas isoladamente, quer quando relacionadas a outras palavras, através do conceito de distância visto neste trabalho, ou através de novas métricas a serem propostas;

- à aplicação de dicionário de sinônimos ou dicionário da área em ambientes como o aqui proposto para favorecer o encontro entre palavras resultantes do tratamento dado pelo indexador com palavras existentes em tesauros;

- as novas métricas advindas de textos científicos. Neste trabalho apurou-se a medida de cada unidade de texto, para todos textos disponíveis na base. Estes dados ficarão para futuros trabalhos, onde seja possível compreender a importância de palavras colocadas em unidades específicas de texto a

partir do universo que esta parte encera e sua importância na representação do conteúdo;

- à busca de outras formas para liberar o profissional de informação cada vez mais para atividades intelectuais, cabendo à máquina o processamento mais pesado e a ele a decisão e a escolha;

- ao entendimento da interdisciplinaridade da Ciência da Informação através da vivência em pesquisas e projetos com outras áreas;

- ao fomento de projetos que disponibilizem, no formato digital, textos científicos de periódicos com publicação anterior ao advento da Web, facilitando métodos para averiguação do grau de novidade ou obsolescência de termos e de ferramentas de indexação;

- aos projetos cuja massa de textos científicos inclua artigos de congressos, cuja velocidade de publicação e a legitimidade da pertinência da área, favoreceram a seleção de termos que traduzam novidade.

Fica a certeza de que apenas iniciou-se um caminho.

REFERÊNCIAS

AITCHISON, Jean; GILCHRIST, Alan. **Manual para construção de tesauros**. Rio de Janeiro: BNG, 1979.

ALVARENGA, Lídia. A teoria do conceito revisitada em conexão com ontologias e metadados no contexto das bibliotecas tradicionais e digitais. **DataGramaZero -** Revista de Ciência da Informação, v.2, n.6, dez. 2001. Disponível em: <http://www.dgzero.org/dez01/F_I_art.htm>. Acesso em: 15 fev. 2005.

ANDERSON, David R.; SWEENEY, Dennis J.; WILLIAMS, Thomas A. **Estatística aplicada à Administração e Economia**. Tradução de Luiz Sérgio de Castro Paiva. São Paulo: Pioneira Thomson Learning, 2002.

ARAÚJO, Vânia M. R. H. Sistemas de recuperação da informação – SRIs. In: ___. **Sistemas de recuperação da informação**: nova abordagem teórico-conceitual. 1994. Tese (Doutorado em Comunicação e Cultura) – Escola de Comunicação, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 1994. cap. 5, p. 84-122.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS – ABNT. **NBR 10520**: informação e documentação: citações em documentos – apresentação. Rio de Janeiro, 2002a.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS – ABNT. **NBR 6022**: informação e documentação: artigo em publicação periódica científica impressa – apresentação. Rio de Janeiro, 2003.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS – ABNT. **NBR 6023**: informação e documentação: elaboração de referências. Rio de Janeiro, 2002.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS - ABNT. **NBR 6024**: Informação e documentação: numeração progressiva das seções de um documento. Rio de Janeiro, 2003.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS - ABNT. **NBR 6028**: resumos. Rio de Janeiro, 2003.

BAEZA-YATES, R.; RIBEIRO NETO, B. **Modern information retrieval.** New York: Addison-Wesley, 1999.

BARRETO, Aldo de Albuquerque. Políticas de monitoramento da informação por compressão semântica dos seus estoques. **DatagramaZero -** Revista de Ciência da Informação, v.4, n.2, abr. 2003. Disponível em: <www.datagramazero.org.br>. Acesso em: 03 jul. 2004.

BIBLIOTECA NACIONAL (Brasil). **Histórico**. **Fundação da Biblioteca Nacional**. Disponível em: <www.bn.br>. Acesso em: 12 dez. 2004.

BITI - Biblioteconomia, Informação e Tecnologia da Informação. **Elaboração de tesauros documentários**: tutorial. Disponível em: <http://www.conexaorio.com/biti >. Acesso em: 06 mar. 2005.

BOUDON, Raymond. **Os métodos em sociologia**. Tradução de Lólio Lourenço de Oliveira. São Paulo: Ática, 1989.

BRAGA, Lílian Maria. **Palavras de títulos e resumos como acesso ao conteúdo do documento**: uma análise numérica. 1982. Dissertação (Mestrado em Ciência da Informação) – Instituto Brasileiro de Informação em Ciência e Tecnologia/Universidade Federal do Rio de Janeiro, Rio de Janeiro, 1982.

BRASIL. Ministério da Ciência e Tecnologia. **Portal Fust - Bibliotecas**. Glossário de biblioteconomia e documentação. Disponível em: <http://portalfust.socinfo.org.br/contribuicoes.htm>.Acesso em: 21 out. 2005.

BRUANDET, M. F. Concept notion for automatic and dynamic thesaurus updating. In: ANNUAL INTERNATIONAL CONFERENCE ON SYSTEMS DOCUMENTATION, 1., 1982, Carson. **Actas...** Carson: ACM Press, 1982. p. 18-22.

BRUSCHINI, Cristina; ARDAILLON, Danielle; UNBEHAUM, Sandra G. **Tesauro para estudos de gênero e sobre mulheres**. 34. ed. São Paulo: Fundação Carlos Chagas, 1998.

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

BUSSAB, Wilson de O., MORETIN, Pedro A. Estatística básica. São Paulo: Saraiva, 2002.

CAMPOS, Maria Luiza de Almeida. **Linguagem documentária**: teorias que fundamentam sua elaboração. Niterói: Ed UFF, 2001.

CAMPOS, Maria Luiza de Almeida. Perspectivas para o estudo da área de representação da informação. **Ciência da Informação**, Brasília, v.25, n.2, 1995.

CASTELLS, Manuel. A revolução da tecnologia da informação. In: _____. **A sociedade em rede**. 2. ed. São Paulo: Paz e Terra, 1999. v.1, p. 49-86.

CAVALCANTI, Cordélia R. **Indexação & tesauro**: metodologia & técnicas. Brasília: Associação de Bibliotecários do Distrito Federal, 1978.

CHARTIER, Roger. **A aventura do livro**: do leitor ao navegador; conversações com Jean Lebrun. São Paulo: Ed. UNESP, 1998.

CHAUMIER, Jacques. **As técnicas documentais**. Lisboa: Europa América, 1971.

CHAUMIER, Jacques. **Les langages documentaires**: le traitement linguistique de l'information documentaire. Paris: Entreprise Moderne, 1978.

CHEN, Kuang-hua; Wu, Chien-tin. Automatically controlled-vocabulary indexing for text retrieval. In: COMPUTATIONAL LINGUISTICS CONFERENCE, 12., 1999, Hsinchu. **Actas...** Hsinchu: [s.n.], 1999, p.171-185.

CHEN, Z. et al. Web: building a web thesaurus from web link structure. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 26., 2003, Toronto. **Actas…** Toronto: ACM Press, 2003. p. 48-55.

CINDOC. Tesauro en Biblioteconomía y Documentación. Disponível em: <http://www.cindoc.csic.es>. Acesso em: 01 jun. 2005.

CINTRA, A. M. M.; TÁLAMO, M. F. G. M.; LARA, M. L. G.; KOBASHI, N. Y. **Para entender as linguagens documentárias**. 2. ed. rev. e ampl. São Paulo: Polis, 2002.

COLLISON, Robert L. **Índices e indexação**. São Paulo: Polígono, 1972.

CONSELHO NACIONAL DE DESENVOLVIMENTO CIENTÍFICO E TECNOLÓGICO – CNPq. História. Disponível em <http://www.cnpq.br/sobrecnpq/historia/>, Acesso 25 out. 2005.

CRAVEN, Tim. **Thesaurus construction**. Faculty of Information and Media Studies University of Western Ontario. Disponível em: <http://instruct.uwo.ca/gplis/677/thesaur/main00.htm>. Acesso em: 18 mar. 2005.

CROUCH, C. J.; YANG, B. Experiments in automatic statistical thesaurus construction. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 15., 1992, Copenhage. **Actas…** Copenhage: ACM Press, 2003. p. 77-88.

CUNHA, H. R. S. **Padrão PUC Minas de normalização**: normas da ABNT para apresentação de artigos em periódicos científicos. Disponível em: <www3.pucminas.br/documentos/normalizacao_artigos.pdf>. Acesso em: 27 fev. 2005.

CURRÁS, Emilia. **Tesauros, linguagens terminológicas**. Brasília: IBICT, 1995.

DAHLBERG, Ingetraut. O futuro das linguagens de indexação. In: CONFERÊNCIA BRASILEIRA DE CLASSIFICAÇÃO BIBLIOGRÁFICA, 1972, Rio de Janeiro. **Anais...** Brasília: IBICT, 1979a. v. 1, p. 323-334.

DAHLBERG, Ingetraut. Teoria da classificação ontem e hoje. In: CONFERÊNCIA BRASILEIRA DE CLASSIFICAÇÃO BIBLIOGRÁFICA, 1972, Rio de Janeiro. **Anais...** Brasília: IBICT, 1979b. v.1, p.352-370.

DAHLBERG, Ingetraut. Teoria do conceito. **Ciência da Informação**, Rio de Janeiro, v.7, n.2, p.101-107, 1978.

DICIONÁRIO universal da língua portuguesa. 3. ed. Lisboa :Texto,1998. Disponível em :< http://www.priberam.pt>. Acesso em: 19 mar. 2005.

DOCUTES. Tesauro de Ciencias de la Documentación. Disponível em: <http://www3.unileon.es/dp/abd/tesauro/pagina/tesdocumentacion/docutes.htm>, Acesso em: 01 jun. 2005.

DODEBEI, Vera Lúcia Doyle. **Tesauro**: linguagem de representação da memória documentária. Niterói: Intertexto, 2002.

DOMINGUES, Ivan. Conhecimento e Transdisciplinaridade II: aspectos metodológicos. Belo Horizonte: Editora UFMG, 2005.

DONGHONG, J.; JUNPING, G.; CHANGNING, H. Combining a chinese thesaurus with a chinese dictionary. In: CONFERENCE ON ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 36, 1998. **Actas…** [s.l.]: [s.n.], 1998. v.1.

DROZD, L. Some remarks on a linguistic theory. In: THEORETICAL AND METHODOLOGICAL PROBLEMS OF TERMINOLOGY. Nov. 27-30, 1979. Moscow. Proceedings ... Muenchen: Saur, 1981. p. 106-117.

EASYPHP. Disponível em: < http://www.easyphp.org/>. Acesso em: 01 jun. 2005.

ELMASRI, R.; NAVATHE, S. B. **Fundamentals of database systems**. 3.ed. Addinson Wesley, 2000.

FERREIRA, Aurélio Buarque de Holanda. **Minidicionário da língua portuguesa.** 3.ed. Rio de Janeiro : Nova Fronteira, 1989. p. 73.

FOO, S. et al. Automatic thesaurus for enhanced Chinese text retrieval. **Library Review**, v.49, n.5, p.230-239, July 2000.

FOSKETT, Anthony Charles. **A abordagem temática da informação**. São Paulo: Polígono, 1973.

FRANÇA, J. L. et al. **Manual para normalização de publicações técnico-científicas**. 7.ed. rev. e aum. Belo Horizonte: UFMG, 2004.

FREIRE, Isa Maria. Da construção do conhecimento científico à responsabilidade social da ciência da informação. **Informação & Sociedade**, João Pessoa, v.12 n.1 2002. Disponível em: <http://www.informacaoesociedade.ufpb.br/pdf/IS1210207.pdf>. Acesso em: 18 ago. 2005.

FREUND, John E. Estatística aplicada: economia, administração e contabilidade. Porto Alegre: Bookman, 2000.

FREUND, John E. Estatística aplicada: economia, administração e contabilidade. Porto Alegre: Bookman, 2000.

FUNDAÇÃO DE AMPARO À PESQUISA DO ESTADO DE SÃO PAULO - Fapesp**. Histórico.** disponível em: <http://www.fapesp.br/materia.php?data[id_materia]=1>. Acesso em: 25 out. 2005.

GILCHRIST, Alan. **The thesaurus in retrieval**. London: Aslib, 1971.

GOMES, Hagar Espanha (Org.). **Manual de elaboração de tesauros monolíngües**. Brasília: Programa Nacional de Bibliotecas de Instituições de Ensino Superior, 1990.

GOMES, Hagar Espanha; CAMPOS; Maria Luiza de Almeida. Tesauro e normalização terminológica: o termo como base para intercâmbio de informações. **DataGramaZero -** Revista de Ciência da Informação, Rio de Janeiro, v.5, n.6, dez. 2004. Disponível em: <http://www.dgzero.org/dez04/Art_02.htm >. Acesso em: 01 jul. 2005.

GUIMARÃES, Reinaldo F. N. Pesquisa no Brasil: a reforma tardia. **São Paulo Perspectiva**, v.16, n.4, p.41-47, out./dez. 2002.

HODGE, Victoria J.; AUSTIN, Jim. Hierarchical word clustering – automatic thesaurus generation. **Neurocomputing,** v.48, n.1-4, p.819-846, Oct. 2002.

HOLANDA, A. et al. Thesaurus as complex network. **Physica, A: Statistical Mechanics and its Applications**, v. 344, iss. 3-4 [SPECIAL ISSUE], p. 530-536, 2004.

INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA E TECNOLOGIA – IBICT. **Tesauro de Ciência da Informação**: versão preliminar. Brasília: IBICT, 1989.

ISOC-DC. Base de Datos de Bibliotecomía, Documentación y Política Científica. Disponível em: <http://bddoc.csic.es:8080/BIBYDOC/BASIS/bibydoc/web/docu/SF>. Acesso em: 04 jul. 2005.

JESUÍNO, Jorge Correa. O método experimental nas ciências sociais. In: SILVA, Augusto Santos; PINTO, José Madureira (Org.). **Metodologia das Ciências Sociais.** 8 ed. Porto: Edições Afrontamento, 1986. p. 215-249.

JONES, K; WILLET, P. **Readings in information retrieval**. California: Morgan Kaufmann, 1997. p.15-20.

JOYCE, T.; NEEDHAM, R. M. The thesaurus approach to information retrieval. **American Documentation**, v.9, p.192-197, 1958.

KAULA, Prithvi N. Repensando os conceitos no estudo da classificação. Tradução dos Editores do Herald of Library Science. **BITI – Biblioteconomia, Informação & Tecnologia da Informação**. Disponível em: <http://www.conexaorio.com/biti/>. Acesso em: 23 fev. 2005. Título original: Rethinking on the concepts in the study of classification.

KIMOTO, H.; IWADERA, T. Construction of a dynamic thesaurus and its use for associated information retrieval. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 13, 1989, Brussels. **Actas...** Brussels: ACM Press, 1989. p.227-240

KOSHIBA, L.; PEREIRA, D. M. F. História do Brasil. 5. ed. rev. e ampl. São Paulo: Atual, 1987. p. 386.

KREMER, Jeannette M. Estratégia de busca. **Revista da Escola de Biblioteconomia da UFMG**, Belo Horizonte, v.14, n.2, p.187-220, set. 1985.

KRZYZANOWSKI, R.F.; FERREIRA, M. C. G. Avaliação de periódicos científicos brasileiros. **Ciência da Informação**, Brasília, v.27, n.2, p.165-169, maio/ago.1998.

KUMAR, Krishan. A sociedade de informação. In: ___. **Da sociedade pós-industrial à pós-moderna:** novas teorias sobre o mundo contemporâneo. Tradução de Ruy Jungmann. Rio de Janeiro: Jorge Zahar, 1997. cap. 2, p. 18-47.

LANCASTER, F. W. **Indexação e resumos**: teoria e prática. Brasília: Briquet de Lemos, 1993.

LANCASTER, F. W.; WARNER, Amy J. **Information retrieval today**. Arlington: Information Resources, 1993.

LAVILLE, Cristian; DIONNE, Jean. **A construção do saber**: manual de metodologia da pesquisa em ciências humanas. Tradução de Heloísa Monteiro e Francisco Settineri. Belo Horizonte: UFMG, 1999.

LEMOS, Antônio Agenor Briquet de. Presente e futuro do periódico científico. **Correio Braziliense**, Brasília, 13 jul. 1968. Caderno Cultural, p. 3. Disponível em: < http://www.briquetdelemos.com.br/editor1.htm>. Acesso em: 20 out. 2005.

LIMA, Gercina Ângela Borém de O. **Mapa Hipertextual (MHTX)**: um modelo para organização hipertextual de documentos. 2004. Tese (Doutorado em Ciência da Informação) – Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2004.

LOPES, Ilza Leite. Uso das linguagens controlada e natural em bases de dados: revisão de literatura. **Ciência da Informação**, Brasília, v.31, n.1, p.60-71, maio/ago. 2002.

LOPES NETO, D. et al. Análise de títulos de artigos de pesquisas publicadas em um periódico brasileiro de enfermagem. **Revista Latino-Americana de Enfermagem,** Ribeirão Preto, v.10, n.1, p.77-84, jan./fev. 2002.

LUHN, H. P. The automatic derivation of information retrieval encodements from machine-readable texts. In: SPARCK JONES, K.; WILLET, P. **Readings in information retrieval**. San Francisco: Morgan Kaufmann, 1997. p.21-24.

MANGUEL, Alberto. **Uma história da leitura**. 2. ed. São Paulo: Companhia das Letras, 2001.

MAPLE, Amanda. **Faceted access**: a review of the literature. Disponível em: <http://www.music.indiana.edu/tech_s/mla/facacc.rev>. Acesso em: 31 mar. 2005.

MARQUES DE JESUS, Jerocir Botelho. Tesauro: um instrumento de representação do conhecimento em sistemas de recuperação da informação. In: SEMINÁRIO NACIONAL DE BIBLIOTECAS UNIVERSITÁRIAS, 12., 2002, Recife. Disponível em: <http://www.ndc.uff.br/textostecnicos.asp> . Acesso em: 03 jul. 2004.

MARTINS, Wilson**. A palavra escrita**: história do livro, da imprensa e da biblioteca. 3.ed. rev. e atual. São Paulo: Ática, 1998.

McGARRY, Kevin. **O contexto dinâmico da informação**: uma análise introdutória. Brasília: Briquet de Lemos, 1999.

MEADOWS, A. J. **A comunicação científica**. Brasília: Briquet de Lemos/Livros, 1999. 268 p.

MILSTEAD, Jéssica L. **ASIS thesaurus of information science and librarianship.** 2. ed. New Jersey: Information Today, 1998.

MOREIRA, Manoel Palhares. MOURA, Maria Aparecida. **Geração automática de tesauros: abordagem conceitual e viabilidade tecnológica**. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 6., 2005, Florianópolis. *Anais eletrônicos*. Florianópolis, 2005. 1 cd-rom.

MOREIRA, Manoel Palhares, STEMPLIUC, Sergio Murilo, PFEILSTICKER, Igor Seufferheld de Oliveira. **Formação de um vocabulário controlado a partir de palavras-chave**. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 6., 2005, Florianópolis. *Anais eletrônicos*. Florianópolis, 2005. 1 cd-rom.

MORITA, K. et al. Word classification and hierarchy using co-occurrence. **Information Processing and Management**, v.40, n.6, p.957-972, Nov. 2004.

MOSTAFA, Solange Puntel. Ciência da Informação: uma ciência, uma revista. **Ciência da Informação,** Brasília, v.25, n.3, p.1-5.,1996.

MOTTA, Dilza Fonseca. **Modelo relacional como nova abordagem para a construção de tesauros**. Disponível em <http://www.conexaorio.com/biti/>. Acesso em: 18 jan. 2005.

MOURA, Maria Aparecida; SILVA, Ana Paula; AMORIM, Valéria Ramos de. A concepção e o uso das linguagens de indexação face às contribuições da Semiótica e da Semiologia. **Revista Informação & Sociedade:** estudos**,** João Pessoa, v.12, n.1, p.1-22, 2002.

MULTISYSTEMS. **Multites**. Disponível em: <http://www.multites.com/>. Acesso em: 03 jul. 2005.

MURRIEL, Gatti. ¿Por qué prestar atención al lenguaje? **Boletín Informativo de Temas Lingüísticos del Departamento Académico de Humanidades de la Universidad del Pacífico**, Lima, v.1, n.1, jul. 1998. Disponível em <http://www.up.edu.pe/coine/Boletin1/TRASFOND.HTM>. Acesso em: 10 jun. 2005.

MYSQL. Disponível em: < http://www.mysql.com/>. Acesso em: 01 jun. 2005.

NAVES, Madalena M. L. **Fatores interferentes no processo de análise de assunto**: estudo de caso de indexadores. 2000. Tese (Doutorado em Ciência da Informação) – Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2000.

NAVES, Madalena. M. L. Análise de assunto: concepções. **Revista de Biblioteconomia de Brasília**, Brasília, v.20, n.2, p.215-226, jul./dez.1996.

OLIVEIRA, Érica B. P. M. de; NORONHA, Daisy P. A comunicação científica e o meio digital. **Informação & Sociedade**: estudos, João Pessoa, v.15, n.1, 2005. Disponível em: <http://www.informacaoesociedade.ufpb.br/>. Acesso em: 25 out. 2005.

PHP. Hypertext Preprocessor. Disponível em:< http://www.php.net.> Acesso em: 01 jun. 2005.

PINHEIRO, Lena Vânia Ribeiro. Campo interdisciplinar da Ciência da Informação: fronteiras remotas e recentes. In: _____ . CASTRO, Ana Lucia Siaines de. (Orgs.). **Ciência da informação, ciências sociais e interdisciplinaridade.** Rio de Janeiro: Instituto Brasileiro de Informação em Ciência e Tecnologia, 1999.

PINKER, Steven. **O instinto da linguagem**: como a mente cria a linguagem. São Paulo: Martins Fontes, 2002.

RANGANATHAN, Shiyali Ramamrita. Prolegomena to library classification. 3. ed. London: 1967.

RICHARDSON, Roberto Jarry et al. **Pesquisa Social**: métodos e técnicas. São Paulo: Atlas, 1999.

ROBERTS, Norman. The pre-history of the information retrieval thesaurus. **Journal of Documentation**, v.40, n.4, p.271-285, Dec. 1984.

RODRIGUES, Marly. **A década de 50**: populismo e metas desenvolvimentistas no Brasil. 3.ed. São Paulo: Ática, 1996.

ROGET, Peter Mark. **Thesaurus of English words and phrases**. [s.l.]: [s.n.], 1960.

SALTON, Gerard. A Blueprint for automatic indexing. **SIGIR Fórum,** v.16, n.2, p. 22-38, May/June 1981.

SALTON, Gerard. Automatic text indexing using complex identifiers. In: ACM CONFERENCE ON DOCUMENT PROCESSING SYSTEMS, 1988, Santa Fé. **Actas** … Santa Fé: ACM Press, 1988, p.135–144.

SALTON, Gerard. **Dynamic information processing**. New Jewrsey, Prentice Hall, 1975.

SALTON, Gerard; McGILL, Michael J. **Introduction to modern information retrieval.** New York: McGraw Hill Book, 1983.

SARACEVIC, T. Ciência da Informação: origem, evolução e relações. **Perspectivas em Ciência da Informação,** v.1, n. 1, (jan/jun.) 1996. Belo Horizonte: Escola de Biblioteconomia da UFMG, 1996.

SARACEVIC, T. Evaluation of evaluation in information retrieval. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 18., 1995. Seattle. **Actas...** Seattle: ACM Press, 1995. p.138-146.

SHERA, Jesse H. Padrão, estrutura e conceituação na classificação. Tradução de Hagar Espanha Gomes. **BITI – Biblioteconomia, Informação & Tecnologia da Informação**. Disponível em: <http://www.conexaorio.com/biti/>. Acesso em: 14 fev. 2005.

SHIRI, A.; REVIE, C. Thesauri on the web: current developments and trends. **Online Information Review**, v.24, n.4, p.273-279, 2000.

SILVA, Augusto Santos; PINTO, José Madureira. Uma visão global sobre as ciências sociais. In:____ (Org.). **Metodologia das Ciências Sociais.** 8 ed. Porto: Edições Afrontamento, 1986. p. 9-27.

SOERGEL, D. **Indexing languages and thesauri: construction and maintenance**. Los Angeles, Cal.: Melville, 1974.

SOUZA, Raimundo Ferreira de. A história das bibliotecas. **UFAC na imprensa**. Disponível em: <http://www.ufac.br//imprensa>. Acesso em: 18 nov. 2004.

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

SPINK, A. et al. Searching the web: the public and their queries. **Journal of the American Society for Information Science and Technology**, v.52, n.3, p.226-234, 2001.

SPITERI, Louise. A simplified model for facet analysis. **Canadian Journal of Information and Library Science**, Toronto, v.23, n.1/2, p.1-30, Apr./Jul. 1998.

SVENONIUS, Elaine. **The intellectual foundation of information organization**. Cambridge, MA: MIT Press, 2000.

TARGINO, Maria das Graças. Comunicação científica na sociedade tecnológica: periódicos eletrônicos em discussão. **Comunicação e Sociedade,** São Bernardo do Campo, n. 31, p. 71-98, 1. sem. 1999.

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

VALÉRIO, Palmira Moriconi. **Espelho da Ciência**: avaliação do Programa Setorial de Publicações em Ciência e Tecnologia da FINEP. Brasília: FINEP/IBICT, 1994.

VICKERY, Brian Campbell. **Classification and indexing in science**. 3.ed. London: Butterworths, 1975.

WEBCHOIR. The Thesaurus Construction System (TCS). Disponível em: <http://www.webchoir.com>. Acesso em: 01 jun. 2005.

WILSON, Patrick. Subjects and the sense of position. In: CHAN, Lois Mai; RICHMOND, Phyllis A.; SVENONIUS, Elaine (Ed.). **Theory of subject analysis**: a sourcebook. Littleton, Colorado: Libraries Unlimited, 1985. p. 306-325.

Anexo 1 Páginas do TCI – Tesauro em Ciência da Informação

Figura 23- TCI: Tela inicial

Figura 24 - TCI: ordem alfabética de termos

Figura 25 - ´TCI: Ordem hierárquica de termos

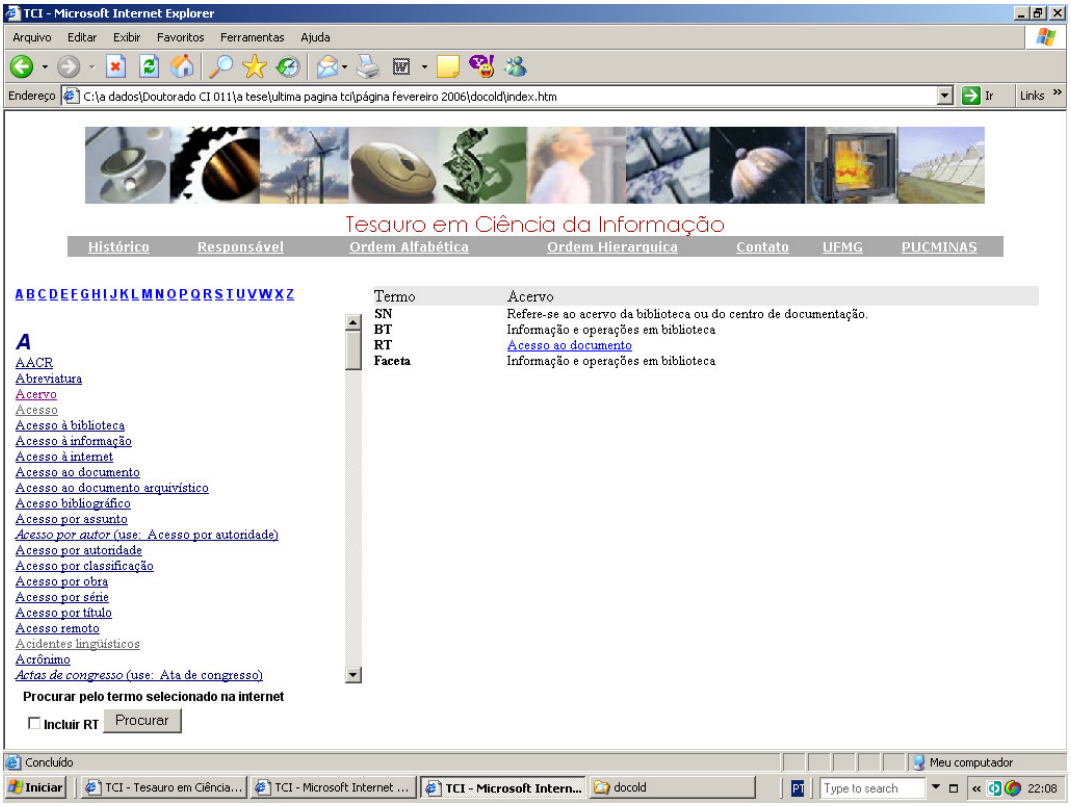

Figura 26 - TCI: apresentação de termos

Anexo 2 Apresentação do cálculo do escore padronizado das freqüências

Para se comparar unidades de medidas de diferentes conjuntos de dados, utiliza-se o escore padronizado. Se ***x*** é uma mensuração pertencente a um conjunto de dados com média $\overline{X}$ (Figura 23) e desvio padrão ***s*** (Figura 24), então seu valor em unidades padronizadas, ***EP***, equivale ao calculo da diferença entre seu valor e o valor médio dividido pelo desvio padrão (FREUND, 2000), conforme apresentado na Figura 25.

$$\overline{X} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Figura 27 - Fórmula para cálculo da média de um série de valores

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - x_m)^2}{(n-1)}}$$

Figura 28 - Fórmula para cálculo do desvio padrão

$$EP = \frac{x - x_m}{s}$$

Figura 29 - Fórmula para cálculo do escore padronizado

O escore padronizado informa a quantos desvios-padrão um valor está acima ou abaixo da média do conjunto de dados, permitindo comparações entre dados de conjuntos distintos.

O valor zero da média dos escores padronizados é explicado pela própria estatística. Dados uma série de *n* observações de uma mesma variável *X*, denotada como $x_1, x_2, x_3, \ldots, x_n$., o cálculo da média dessas observações é obtido pela fórmula da Figura 23. O escore padronizado é calculado de acordo com Figura 25.

A seguir, transforma-se a série de observações $x_1, x_2, x_3, \ldots, x_n$ em uma nova série de observações denotada como $z_1, z_2, z_3, \ldots, z_n$.

O cálculo da média e do desvio padrão dessa nova variável Z, que é uma transformação de X ficaria de acordo com a Figura 26, ou seja zero.

$$\overline{Z} = \frac{z_1 + z_2 + \cdots + z_n}{n} = \frac{EP(x_1) + EP(x_2) + \ldots + EP(x_n)}{n} = \frac{\frac{(x_1 - \overline{X})}{s_x} + \frac{(x_2 - \overline{X})}{s_x} + \ldots + \frac{(x_2 - \overline{X})}{s_x}}{n} =$$

$$= \frac{x_1 - \overline{X} + x_2 - \overline{X} + \ldots + x_n - \overline{X}}{n.s_x} = \frac{(x_1 + x_2 + \cdots + x_n) - n.\overline{X}}{n.s_x} = \frac{x_1 + x_2 + \cdots + x_n}{n.s_x} - \frac{n.\overline{X}}{n.s_x} =$$

$$= \frac{1}{s_x}.\frac{x_1 + x_2 + \cdots + x_n}{n} - \frac{1}{s_x}.\overline{X} = \frac{1}{s_x}.\overline{X} - \frac{1}{s_x}.\overline{X} = 0$$

Figura 30 - Calculo da média dos escores padronizados

Ou seja, a média de uma série de escores padronizados é sempre igual a zero. Seguindo o mesmo raciocínio, porém de forma mais complexa, podemos demonstrar que o desvio padrão de uma série de escores padronizados é sempre igual a 1.

Sempre que se transforma uma série de observações de uma variável em escores padronizados, faz-se a justamente a transformação da escala original para uma escala padrão, na qual a média e o desvio padrão são fixos e conhecidos.

Na apresentação dos resultados foram adotados percentis. Um percentil é uma medida da posição relativa de uma unidade que está sendo observada em relação a todas as outras. O *p*-ésimo percentil tem no mínimo *p*% dos valores abaixo daquele ponto e no mínimo (100 - *p*)% dos valores acima (ANDERSON et. al, 2002).

Anexo 3 Palavras Tratadas

Tabela 26 - Palavras tratadas existentes no TCI

Palavras tratadas existentes no TCI			
Palavra Tratada	Índice	Palavra Tratada	Índice
Informação	11,3331	Administração de Empresas	-0,1998
Conhecimento	4,9986	Protocolo	-0,2017
Ciência da informação	4,0476	Comércio eletrônico	-0,2027
Pesquisa	3,7695	Sistema de recuperação de informação	-0,2042
Sistema	3,5438	Centro de informação	-0,2043
Acesso à informação	3,3655	Cienciometria	-0,2116
Comunicação	3,1056	Máquina de busca	-0,213
Dado	2,5432	Desenvolvimento de coleção	-0,2146
Internet	2,3442	Competência informacional	-0,2147
Rede	2,1917	Classificação facetada	-0,2173
Conceito	2,1201	Indústria da Informação	-0,2191
Modelo	1,9673	Arquivologia	-0,2204
Web	1,7032	Diretório	-0,221
Usuário	1,6824	Biblioteca pública	-0,2224
Biblioteca	1,5709	Indexador	-0,2234
Educação	1,4928	Agente inteligente	-0,2235
Tecnologia da informação	1,3999	SGML	-0,2241
Sociedade da informação	1,2568	Pré-impressão	-0,2284
HTML	1,1063	Web Semântica	-0,2304
Biblioteconomia (área de conhecimento)	1,0307	Página web	-0,2311
Periódico	0,9972	Linguagem de marcação	-0,233
Sistema de informação	0,9948	Lei de Lotka	-0,2373
Realidade virtual	0,9883	Informetria	-0,2396
Termo	0,9356	Sistema de informação gerencial	-0,2417
Base de dados	0,9206	Informação e sociedade	-0,2478
Representação da informação	0,9128	Indexação automática	-0,2504
Arquivo	0,8887	Coleção	-0,2548
Documentação	0,873	Serviço de referência	-0,2548
Publicação	0,7274	Explosão informacional	-0,2559
Tecnologia da informação e da comunicação	0,6759	Desenvolvimento de sistema	-0,2576
Classificação	0,6303	Homografia	-0,261
Pesquisador	0,6286	História natural	-0,2652
Professor	0,6118	Engenharia	-0,2654
Tese	0,6028	MARC	-0,2654
Profissional de informação	0,6023	Museu de ciência	-0,2654
Recuperação de informação	0,5479	Recurso eletrônico	-0,2654
Memória	0,4608	Lei de Bradford	-0,2671
Hipertexto	0,4043	História	-0,2689
Bibliotecário	0,4015	Interação homem máquina	-0,269
Biblioteca digital	0,379	Biblioteca híbrida	-0,2724
Biblioteca universitária	0,3466	Lei de Zipf	-0,2733
Indexação	0,3149	Universidade	-0,2749
Fonte de informação	0,3087	Processamento de linguagem natural	-0,2758
Serviço de informação	0,2875	Literatura cinzenta	-0,2758

Lingüística	0,267		Linguagem natural	-0,2758
Terminologia	0,2562		Biblioteca escolar	-0,2758
Tomada de decisão	0,2384		Texto eletrônico	-0,2759
Cd-rom	0,2028		Semântica	-0,2767
Transferência de informação	0,1882		Fluxo de Informação	-0,2847
Uso da informação	0,1789		Modelo cognitivo	-0,2847
Dissertação	0,1584		Rede neural	-0,2872
Ontologia	0,1381		Descritor	-0,2908
Disseminação da informação	0,1079		Serviço arquivístico	-0,2908
Informação técnico científica	0,1073		Imagem	-0,2908
Software	0,1027		Método estatístico	-0,2934
Psicologia	0,067		Museu de Arte	-0,2952
Tesauro	0,0656		Profissional da informação	-0,2952
Metadado	0,0618		Necessidade de informação	-0,2952
Bibliografia	0,0553		Tecnologia educacional	-0,2959
Informação Científica	0,0512		Bibliotecário de referência	-0,2978
Busca de informação	0,036		Leitor	-0,3003
Bibliometria	0,0301		Revisão de literatura	-0,3021
Medicina	0,0285		Informação gerencial	-0,3021
Informação tecnológica	0,0067		Aquisição	-0,3065
Banco de dados	-0,0127		Documento arquivístico	-0,3065
Economia da informação	-0,0209		Formato bibliográfico	-0,3065
XML	-0,0332		Gestão de documentos	-0,3065
Organização da informação	-0,0417		Acervo	-0,3065
Correio eletrônico	-0,0546		Informação organizacional	-0,3109
Ciências Humanas	-0,0583		Gerencia de coleção	-0,3109
Publicação eletrônica	-0,0919		Arquitetura de sistema	-0,3109
Recurso de informação	-0,0923		Sociologia da informação	-0,3109
Estratégia de busca	-0,0924		Lógica difusa	-0,3109
Linguagem de indexação	-0,1013		Base de dados bibliográficos	-0,3109
Epistemologia	-0,1107		Informação institucional	-0,3109
Intranet	-0,1138		Índice de citação	-0,3109
Hipermídia	-0,1279		Linguagem de programação	-0,3109
Semiótica	-0,1317		Modelo Booleano	-0,3109
Estudo de usuário	-0,1332		Teatro	-0,3152
Base de conhecimento	-0,1362		Administração hospitalar	-0,3152
Informação estatística	-0,1462		Perfil de usuário	-0,3152
Processamento da informação	-0,1468		Análise de citação	-0,3152
Economia	-0,1516		Liberdade de Informação	-0,3152
Documento digital	-0,1543		Medicina veterinária	-0,3152
Dublin Core	-0,1662		Recuperação da informação	-0,3152
Análise documental	-0,1724		Projeto de pesquisa	-0,3152
Sistema especialista	-0,1743		Resumo	-0,3152
Cibernética	-0,1854		Uso de informação	-0,3152
Arquivo aberto	-0,1928		Ciências sociais	-0,3152
Arquivista	-0,1934			

Tabela 27 - Palavras tratadas não existentes no TCI

Palavras tratadas não existentes no TCI				
Palavra tratada	Índice		Palavra Tratada	Índice
Formação	11,5881		VIRTUS	-0,2715
Ciência	5,8183		Contexto da informação	-0,2722
Brasil	4,7234		Interação universidade-empresa	-0,2724
Tecnologia	3,8572		Imagem digital	-0,2733
Texto	3,4272		Modelagem de Dados	-0,2733
Sociedade	3,3662		Teoria do conhecimento	-0,2733
Trabalho	2,9625		Lógica Cultural	-0,2739
Organização	2,0427		Avaliação (software)	-0,2747
Programa	1,8784		Política de comunicação	-0,2747
Estado	1,7347		Reciclagem papel	-0,2751
Poder	1,6775		E-prints	-0,2758
Avaliação	1,5813		Informação cidadania	-0,2758
Documento	1,5479		Controle social	-0,2759
Qualidade	1,4804		Compressão	-0,2767
Jornalismo	1,4552		Semiose	-0,2767
Criação	1,2577		Ciência clássica	-0,2777
Instituto Brasileiro de Informação em Ciência e Tecnologia	1,1735		Direito de propriedade	-0,2777
Estratégia	1,1442		Estratégia metodológica	-0,2777
Fonte	1,0885		Fenômeno da informação	-0,2777
Linguagem	1,0646		Neoliberalismo	-0,2777
Informática	1,0622		Prática profissional	-0,2777
Comunidade	1,0550		Ciência e Tecnologia	-0,2802
Saber	1,0500		Economia Digital	-0,2802
Paradigma	1,0289		Gestão da qualidade	-0,2802
Referência	1,0068		Informação agropecuária	-0,2802
Corpora	0,9476		Rizoma	-0,2802
Interatividade	0,9406		Universidade Federal do Rio Grande do Norte	-0,2802
Ensino	0,8992		Processos de Informação	-0,2803
Definição	0,8279		Antropologia Brasileira	-0,2811
Horizon	0,7751		Classificação industrial	-0,2811
Competência	0,7551		Controle bibliográfico institucional	-0,2811
Inovação	0,6696		Iluminismo	-0,2811
Disseminação	0,6680		Informação sobre Risco	-0,2811
Liber	0,6429		Normalização Terminológica	-0,2811
Decisão	0,6299		Pesquisa científica	-0,2811
Leitura	0,6072		Qualidade (serviço de biblioteca)	-0,2811
Escrita	0,5977		Consulta pública	-0,2821
Complexidade	0,5683		Sistema de inteligência competitiva	-0,2821
Governo	0,5601		Reversibilidade	-0,2828
Mpress	0,5500		Budapest open access initiative	-0,2846
Padrão	0,4795		Cultura impressa	-0,2846
Produção científica	0,4559		Programa 5S	-0,2846
Globalização	0,4082		Programa de qualidade	-0,2846
Inteligência competitiva	0,4058		SERVQUAL	-0,2846
Gestão do conhecimento	0,3817		Sociedade aprendente	-0,2846
Gestão da informação	0,3547		Transgênicos	-0,2846
Descrição	0,3489		Ambiente de informação	-0,2855

Cidadania	0,2711		Análise setorial	-0,2855
Competitividade	0,2268		Informação potencializada	-0,2855
França	0,2114		Mensuração Estatística	-0,2855
Identidade	0,2028		Progresso Globalizado	-0,2855
Produção do conhecimento	0,1968		Ajuda técnica	-0,2864
Comunicação científica	0,1688		Criação da informação	-0,2864
Política informacional	0,1615		Imagem em movimento	-0,2864
Gerência	0,1607		Monitoramento de informação	-0,2864
Mediação	0,1523		Popularização da Ciência	-0,2864
Ciberespaço	0,1483		Sistema de informação estatística	-0,2864
Automação	0,1240		Texto Linear	-0,2864
Acessibilidade	0,1101		Acesso à arquivo público	-0,2889
Sociedade do conhecimento	0,0958		Busca	-0,2896
Inclusão	0,0927		Comunicação política	-0,2896
Intercâmbio	0,0855		Disseminação da Pesquisa	-0,2896
Autonomia	0,0782		Informação ambiental	-0,2896
Metáfora	0,0497		Centro de Competência	-0,2908
Cenário	0,0317		Consciência individual	-0,2908
Energia	0,0232		Conteúdo semântico	-0,2908
Periódico científico	0,0219		Controle bibliográfico nacional	-0,2908
Capitalismo	0,0209		Cultura digital	-0,2908
Totalidade	0,0167		Integração de Informações	-0,2908
Meios de comunicação	0,0074		Pessoa portadora de necessidades especiais	-0,2908
Exclusão social	0,0067		Política da informação	-0,2908
Inovação tecnológica	0,0051		Processo legislativo	-0,2908
Política pública	-0,0055		Reserva Técnica	-0,2908
Liderança	-0,0073		Tecnologias da comunicação e informação	-0,2908
Alfabetização	-0,0114		Viabilidade do produto	-0,2908
Ensino a distância	-0,0188		Viabilidade do serviço de informação	-0,2908
Interdisciplinaridade	-0,0227		Crescimento do Conhecimento	-0,2915
Informação negócio	-0,0304		Informação legislativa	-0,2915
Contrato	-0,0327		Campo de Conhecimento	-0,2934
País em desenvolvimento	-0,0354		Democratização da informação	-0,2934
Abstração	-0,0409		Manipulação da informação	-0,2934
Informatização	-0,0424		Atuação profissional	-0,2940
Capital intelectual	-0,0440		Perfil profissional	-0,2940
Transferência de Tecnologia	-0,0571		Aparato Informacional	-0,2952
Propriedade intelectual	-0,0598		Cibernose	-0,2952
Legislação	-0,0652		Cognóscio	-0,2952
Língua portuguesa	-0,0668		Comunidade discursiva	-0,2952
Formação profissional	-0,0700		Documento Eletrônico de Arquivo	-0,2952
Mercado de trabalho	-0,0700		Estoque Informacional	-0,2952
Responsabilidade social	-0,0705		Imagem técnica	-0,2952
Biblioteca virtual	-0,0707		Índice da sociedade da informação (ISI)	-0,2952
Catalogação	-0,0712		Informatose	-0,2952
Tratamento da informação	-0,0781		Literatura branca	-0,2952
cognição	-0,0832		Math-net	-0,2952
Organização do conhecimento	-0,0838		Método de Solução de Problemas	-0,2952
ISIS	-0,0950		Midiologia	-0,2952
Palavras-chave	-0,0979		Monitoramento fontes de informação	-0,2952
Cultura organizacional	-0,0980		Normoterapia	-0,2952
Representação do conhecimento	-0,0983		Paratexto	-0,2952

Bem e serviço	-0,1023		Sistema de informação estratégica	-0,2952
Informação estratégica	-0,1046		Sociologia da Cultura	-0,2952
Interoperabilidade	-0,1061		Valência Sintático-Semântica	-0,2952
Inteligência empresarial	-0,1064		Alinhamento estratégico	-0,2959
Planejamento estratégico	-0,1076		Enfoque sistêmico	-0,2959
Caos	-0,1099		Habilidade informacional	-0,2959
Periódico eletrônico	-0,1115		Cultura Corporativa	-0,2978
Pós-modernidade	-0,1210		Descoberta de conhecimento	-0,2978
Automação de biblioteca	-0,1227		Exclusão digital	-0,2978
Divulgação científica	-0,1251		Financiadora de Estudos e Projetos (Finep)	-0,2978
Valor da informação	-0,1264		Interface de Busca	-0,2978
Ambigüidade (recuperação)	-0,1274		Método de ensino	-0,2978
Dialética	-0,1276		Necessidade de usuário	-0,2978
Ferramenta de busca	-0,1325		Ontologia formal	-0,2978
Modelo conceitual	-0,1346		Orientação a Objetos	-0,2978
Governo Eletrônico	-0,1349		Texto Científico	-0,2978
Universidade Federal do Rio Grande do Sul	-0,1356		Arquivística	-0,2995
Grupo de pesquisa	-0,1403		Corporações transnacionais	-0,2995
Data mining	-0,1456		Estados Unidos	-0,2995
Informação digital	-0,1479		Interpretação legal	-0,2995
Filtro	-0,1482		Metodologia	-0,2995
Estoques de Informação	-0,1492		Organização não-governamental (ONG)	-0,2995
Sociabilidade	-0,1499		Políticas de Formação de Professores	-0,2995
Inteligência organizacional	-0,1502		Publicação automatizada	-0,2995
Hegemonia	-0,1514		Administração de biblioteca	-0,3003
Teoria do conceito	-0,1566		Central de atendimento	-0,3003
Conhecimento tácito	-0,1586		Comportamento de usuários	-0,3003
Comunicação eletrônica	-0,1653		Comunidade científica	-0,3003
Movimento social	-0,1655		Educação tecnológica	-0,3003
Usabilidade	-0,1655		Inovação conservadora	-0,3003
W3C	-0,1710		Produtor rural	-0,3003
Oralidade	-0,1736		Serviço de atendimento ao cliente	-0,3003
Informação eletrônica	-0,1736		Voluntariado	-0,3003
Lista de discussão	-0,1737		Banco de teses	-0,3021
Categorização	-0,1754		Cooperação interinstitucional	-0,3021
Governança	-0,1767		Holismo	-0,3021
Amazônia	-0,1776		Recepção de informação	-0,3021
Aprendizagem organizacional	-0,1787		Rede organizacional	-0,3021
Instituição	-0,1830		Técnica gerencial	-0,3021
Mundo do Trabalho	-0,1848		Treinamento (usuário)	-0,3021
Rede social	-0,1856		Informação estratégia competitiva	-0,3046
Monitoramento ambiental	-0,1864		Institucionalização da informação	-0,3046
Formação de professor	-0,1873		Sociedade de aprendizagem	-0,3046
Copyright	-0,1924		Alfabetização informacional	-0,3065
Auto-organização	-0,1945		Digitalização da informação	-0,3065
Fatores críticos de sucesso	-0,1987		Disseminação Científica	-0,3065
Espaço digital	-0,2008		Ecologia do conhecimento	-0,3065
Information literacy	-0,2029		Gestão participativa	-0,3065
Identidade cultural	-0,2047		Informação para negócios	-0,3065
Desenvolvimento sustentável	-0,2049		Inteligência científica	-0,3065
Lixo	-0,2049		Interface de Consulta	-0,3065
Esquecimento	-0,2050		Modelo Vetorial	-0,3065

Hiperdocumento	-0,2055		Paradigma Informacional	-0,3065
Economia do conhecimento	-0,2068		Política institucional	-0,3065
Conhecimento organizacional	-0,2122		Repositório digital	-0,3065
Usuário da informação	-0,2130		Sinergética	-0,3065
Contrato Social	-0,2136		Sistema de informação para negócios	-0,3065
Direito à informação	-0,2146		Sistema de leitura de tela	-0,3065
Assimilação da informação	-0,2155		Sistema de publicação	-0,3065
Informação Governamental	-0,2186		Teoria da ciência da informação	-0,3065
Processos organizacionais	-0,2187		Análise de referência	-0,3109
Função social	-0,2224		Autoritarismo científico	-0,3109
Biblioteca eletrônica	-0,2241		Avaliação de pares	-0,3109
Terceiro Setor	-0,2241		Cliente-Finep	-0,3109
Critérios de avaliação	-0,2252		Código fonte	-0,3109
Sistema de classificação	-0,2261		Compartilhamento	-0,3109
Jornalismo científico	-0,2267		Compreensão Ativa	-0,3109
Serviço de biblioteca	-0,2277		Diáspora digital	-0,3109
Indicadores bibliométricos	-0,2292		Esoterismo científico	-0,3109
Estado brasileiro	-0,2302		Ética da Informação	-0,3109
Arquivo de log	-0,2304		Forma textual	-0,3109
Gestão	-0,2304		Formação de opinião	-0,3109
Preservação digital	-0,2304		Fundamento social da informação	-0,3109
Economia Política	-0,2309		Gestão de biblioteca	-0,3109
Informação financeira	-0,2311		Gestão de informação	-0,3109
Web site	-0,2311		Gestão Eletrônica de Documento	-0,3109
Governabilidade	-0,2316		Hierarquia da informação	-0,3109
Plataforma Lattes	-0,2321		Holografia	-0,3109
Atividade de informação	-0,2322		Impacto das Tecnologias da Informação	-0,3109
Arquitetura da informação	-0,2330		Índice de Desenvolvimento Humano (IDH)	-0,3109
Conhecimento explícito	-0,2341		Inferência bayesiana	-0,3109
Educação ambiental	-0,2345		Info-exclusão	-0,3109
Ciência cognitiva	-0,2348		Informação ação estratégica	-0,3109
Rede informal	-0,2348		Informação de patente	-0,3109
Humanismo	-0,2353		Informática educacional	-0,3109
Armazenamento	-0,2367		Integração de conhecimento	-0,3109
Mecanismo de busca	-0,2367		Marginalidade Tecnológica	-0,3109
recuperação	-0,2367		Modelo de preservação OAIS	-0,3109
Método científico	-0,2375		Modelo semântico	-0,3109
Memória Social	-0,2384		Neurose virtual	-0,3109
Privatização	-0,2391		Prospecção Informacional	-0,3109
Registro do Conhecimento	-0,2399		Qualidade de informação	-0,3109
Correios	-0,2409		Recurso lingüístico	-0,3109
Produção intelectual	-0,2409		Rede SCienTI	-0,3109
Tecnologia de informação e comunicação	-0,2410		Referência digital	-0,3109
Administração de recursos humanos	-0,2417		Reforma Curricular	-0,3109
Mnemotécnica	-0,2442		Relacionamento com cliente	-0,3109
Apoio à decisão	-0,2453		Saúde	-0,3109
Webmuseu	-0,2453		Serviço de aquisição	-0,3109
Sistema estatístico	-0,2459		Setor informacional	-0,3109
Conteúdo digital	-0,2461		Sistema centrado no usuário	-0,3109
Cultura Informacional	-0,2466		Site jurídico	-0,3109
Construção do Objeto	-0,2468		Sociedade contemporânea	-0,3109
Agente do conhecimento	-0,2470		Sugestão de aquisição	-0,3109

Estatística oficial	-0,2470		Trabalhador rural	-0,3109
Setor industrial	-0,2471		Análise textual	-0,3152
Capacitação de recursos humanos	-0,2479		Aprendizagem	-0,3152
Integração de sistemas	-0,2479		Artigo científico	-0,3152
Filosofia da ciência	-0,2496		Aspecto econômico	-0,3152
Política de pesquisa	-0,2504		Aspecto tecnológico	-0,3152
Transdisciplinaridade	-0,2504		Autoria na Internet	-0,3152
Biologia	-0,2504		Avaliação (periódico)	-0,3152
Serviço Web	-0,2504		Avaliação centrada no usuário	-0,3152
Sintagma nominal	-0,2504		Avaliação de acesso	-0,3152
Webometria	-0,2504		Avaliação impacto tecnológico	-0,3152
Pesquisa operacional	-0,2523		Biblioteca acadêmica	-0,3152
Tempo livre	-0,2530		Cadeia de valor	-0,3152
Informação competitividade	-0,2536		Capacitação (usuário)	-0,3152
Qualidade (serviço)	-0,2539		Catalogação de recursos eletrônicos	-0,3152
Analfabetismo	-0,2540		Classificação do Conhecimento	-0,3152
Conhecimento objetivo	-0,2546		Classificação em Ciência e Tecnologia	-0,3152
Câmara Legislativa do Distrito Federal	-0,2548		CNPq - fomento à pesquisa em Ciência da Informação	-0,3152
Etnografia	-0,2548		Colaboração	-0,3152
ISO 9001	-0,2548		Comportamento informacional - professores educação básica	-0,3152
Leitura Documentária	-0,2548		Curso de graduação	-0,3152
Medicina tropical	-0,2548		Desajuste Conceitual	-0,3152
Núcleo Temático da Seca	-0,2548		Direito da Informação	-0,3152
Servidor de enlace	-0,2548		Disciplina científica	-0,3152
Texto fílmico	-0,2548		Disseminação da Obra Digital	-0,3152
Produtos informacionais	-0,2559		Educação superior	-0,3152
Satisfação do usuário	-0,2559		Elaboração de tesauro	-0,3152
Tipo de usuário	-0,2559		Enlaces	-0,3152
Inclusão digital	-0,2565		Estatística	-0,3152
Pequena e média empresa	-0,2566		Expansão das Novas Tecnologias	-0,3152
Agregado de Informação	-0,2567		Fomento à pesquisa	-0,3152
Auto-arquivamento	-0,2567		Formação continuada	-0,3152
Imagem de síntese	-0,2567		Fundamento	-0,3152
Lingüística estrutural	-0,2567		Gestor da informação	-0,3152
Informação jurídica	-0,2574		Giddens	-0,3152
Monitoramento informação	-0,2574		Indicador empresarial	-0,3152
Paradigma tecnológico	-0,2576		Informação contexto social	-0,3152
Acesso Livre	-0,2584		Informação extradiscursiva	-0,3152
Indução	-0,2585		Informação intradiscursiva	-0,3152
Publicação impressa	-0,2585		Informação mudança social	-0,3152
Sociologia da Ciência	-0,2585		Informação sociedade contemporânea	-0,3152
Função da Terminologia	-0,2599		Informática aplicada	-0,3152
Análise bibliométrica	-0,2601		Inteligibilidade do código fonte	-0,3152
Informação Arquivística	-0,2601		Jornalismo on-line	-0,3152
Comutação bibliográfica	-0,2602		Lilacs	-0,3152
Análise de log	-0,2610		Livro verde	-0,3152
Musealização	-0,2610		Medline	-0,3152
Pesquisa colaborativa	-0,2610		Método quantitativo (avaliação)	-0,3152
Teoria da terminologia	-0,2620		Metodologia curso	-0,3152
Data warehousing	-0,2635		Micro e pequena empresa	-0,3152
Mudança social	-0,2646		Moçambique	-0,3152

Pesquisa participante	-0,2652		Motivação (psicologia)	-0,3152
Programa Sociedade da Informação	-0,2652		Periódico arquitetura e urbanismo	-0,3152
Entidade de classe	-0,2654		Periódico design	-0,3152
Espaço Prisional	-0,2654		Periódico paisagismo	-0,3152
Gestão de investimentos	-0,2654		Pesquisa desenvolvimento	-0,3152
Normose informacional	-0,2654		Planejamento informação	-0,3152
Portal corporativo	-0,2654		Portugal	-0,3152
Sistema de informação de clientes	-0,2654		Produção bibliográfica	-0,3152
Sistema interorganizacional	-0,2654		Produtividade científica	-0,3152
Tesauro Eletrônico	-0,2654		Programa educativo	-0,3152
Massa Documental	-0,2661		Programação código aberto	-0,3152
Regime de informação	-0,2664		Projeto moral e informação	-0,3152
Informação documentária	-0,2671		Publicação bibliográfica	-0,3152
Qualificação profissional	-0,2671		Reflexividade	-0,3152
Ocultamento de Informação	-0,2680		Servidor de link	-0,3152
Setor de informações	-0,2680		Sociedade de risco	-0,3152
Interação homem-computador	-0,2690		Tecnologia hipermídia	-0,3152
Interação ser humano computador	-0,2690		Unidade de conhecimento	-0,3152
Indicador em ciência e tecnologia	-0,2696		Unidade de informação	-0,3152
Sistema estatístico nacional	-0,2696		Usuário de base de dados	-0,3152
Viagem	-0,2697			
Forma simbólica	-0,2715			

Anexo 4 Telas do ambiente AlfaThes

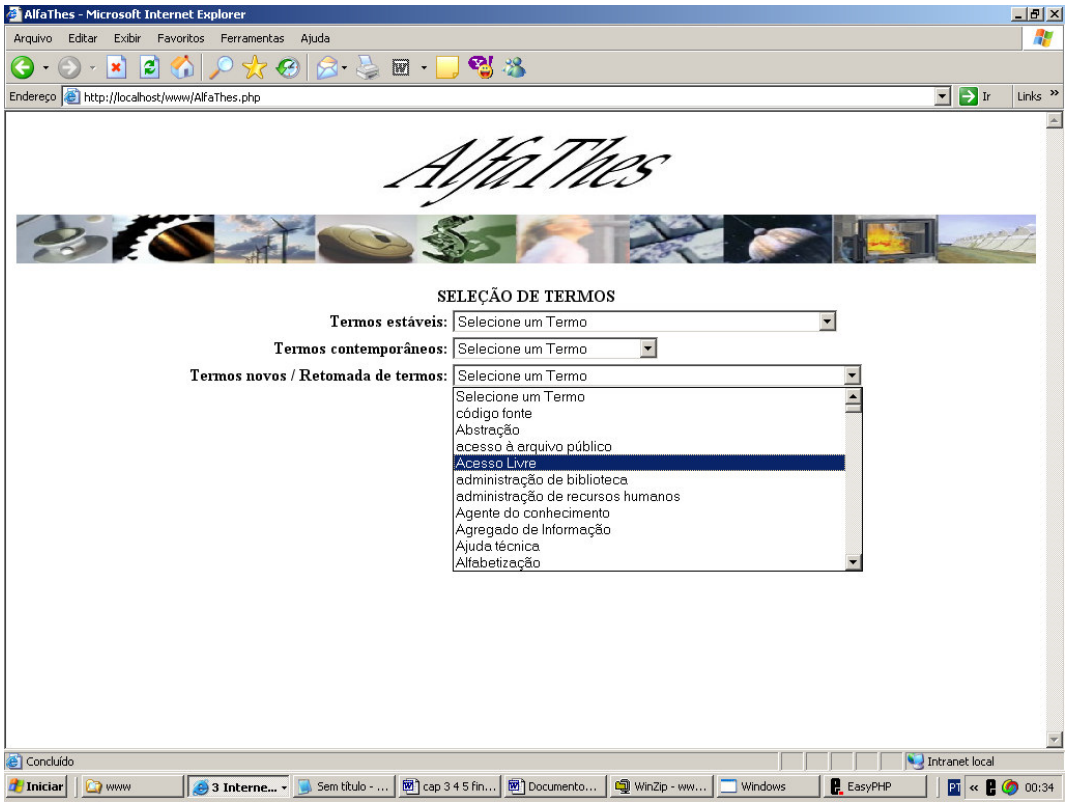

Figura 31 - Ambiente Alfathes: Seleção de termos

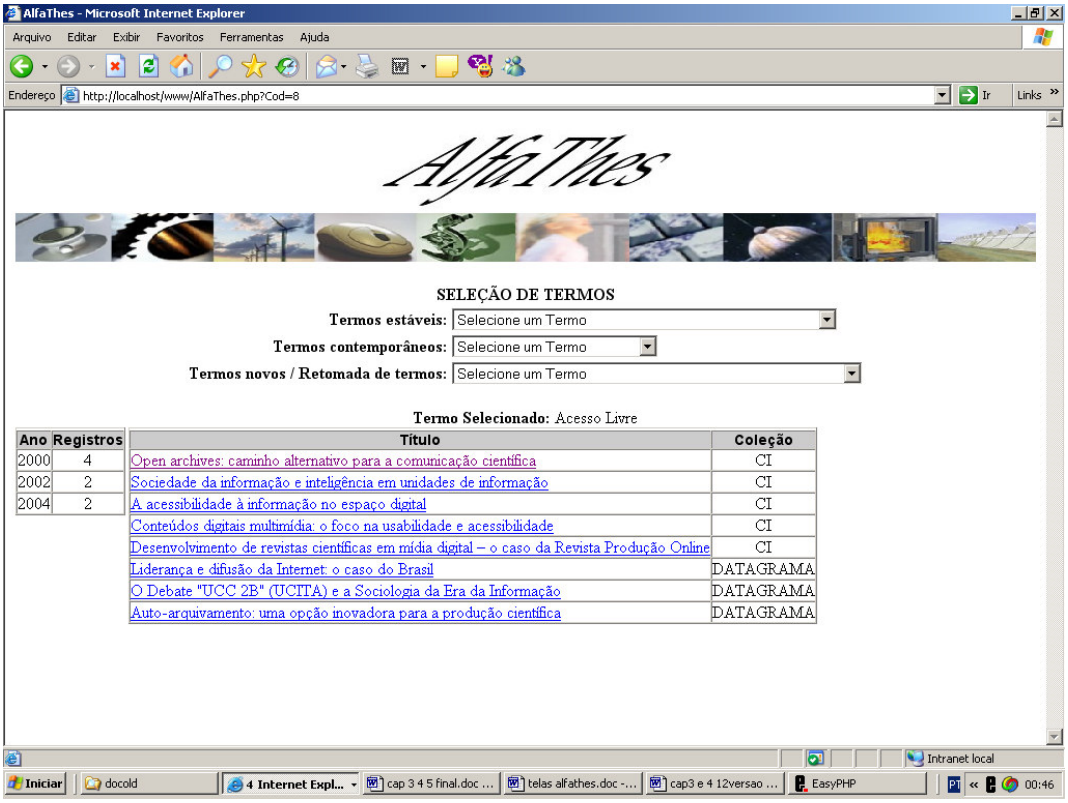

Figura 32 - Ambiente Alfathes: Artigos com termo selecionado

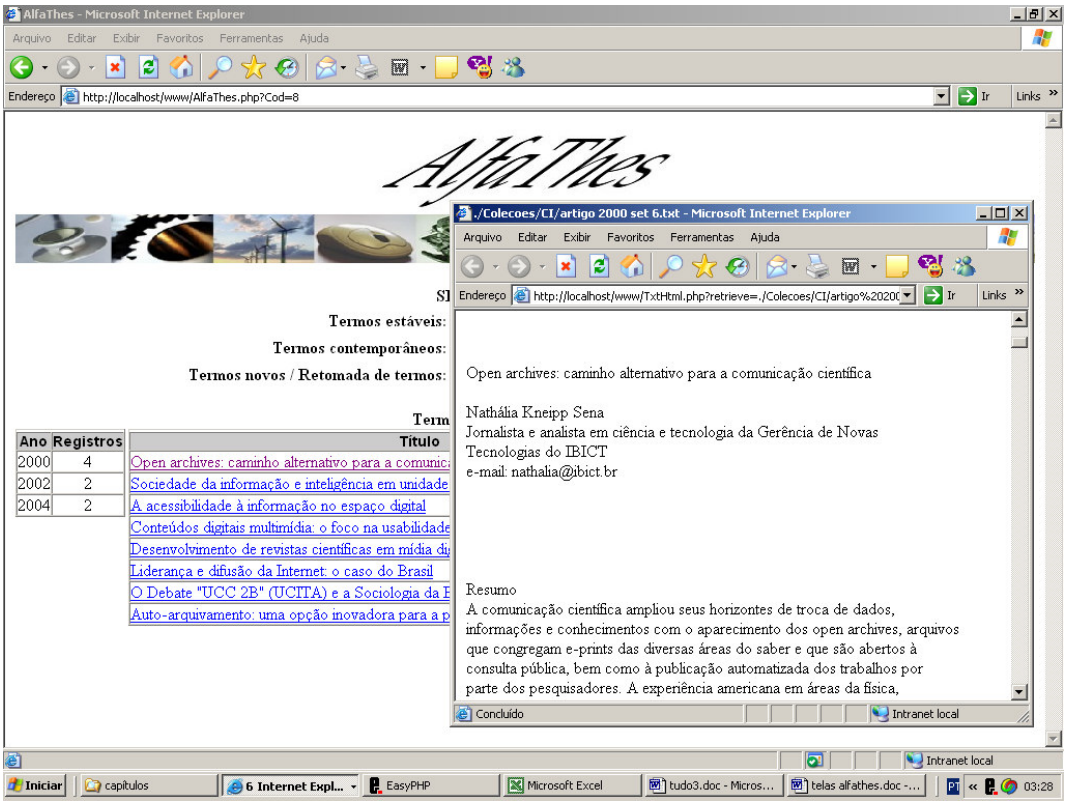

Figura 33 - Ambiente Alfathes: Artigo selecionado

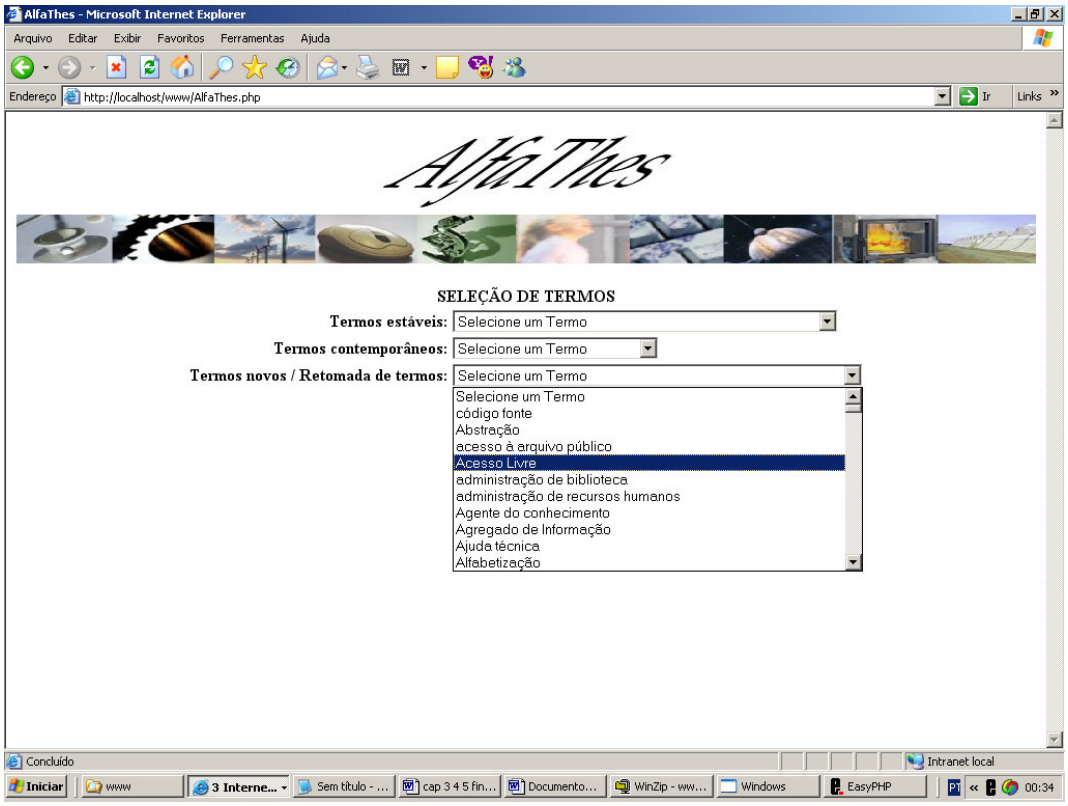

Figura 34 - Ambiente Alfathes: Tela de seleção de termos