

MeSH: de cabeçalho de assunto a tesouro

Eliane Colepícolo¹, Adriano de Jesus Holanda², Evandro Eduardo Seron Ruiz³,
Jacques Wainer⁴, Ivan Torres Pisa⁵

^{1,4,5} Departamento de Informática em Saúde (DIS), Universidade Federal de São Paulo (UNIFESP), Brasil

^{2,3} Departamento de Física e Matemática, Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Universidade de São Paulo (USP), Brasil

Resumo – O estudo e uso das terminologias vêm se tornando cada vez mais essenciais ao desenvolvimento de diversas áreas de pesquisa, incluindo informática em saúde, ciências da informação, inteligência artificial e lingüística computacional. Terminologias tais como vocabulários controlados, cabeçalhos de assunto e tesouros são usados como instrumento para indexação, classificação, busca e recuperação de informação, sendo o tesouro o tipo mais sofisticado e utilizado. O MeSH representa um cabeçalho de assunto das ciências da saúde, publicado e mantido pela National Library of Medicine (NLM), EUA, amplamente utilizado para tratamento de informação em saúde e em ferramentas de aplicação da informática em saúde. O objetivo deste trabalho é apresentar um método de conversão do cabeçalho MeSH em um tesouro para otimizar operações complexas efetuadas com o MeSH em pesquisas científicas. Para exemplificar foi utilizada uma pesquisa sobre a epistemologia da informática em saúde, que visa compreender se a informática em saúde se caracteriza mais como ciência, tecnologia ou uma mistura de ambos, a partir da literatura científica, do uso do MeSH e de métodos e ferramentas computacionais.

Palavras-chave: Informática Médica, Informação em Saúde, Tesouros, Terminologias, Mineração de Textos.

Abstract – The study and use of the terminologies are essential to the development of several research fields, including medical informatics, information sciences, artificial intelligence, and computational linguistic. Terminologies such as controlled vocabularies, subject headers and thesaurus are used as instrument for indexation, classification, search and information retrieval, being the thesaurus the most sophisticated type and used. MeSH is subject header of the life sciences, published and maintained by National Library of Medicine (NLM), USA, largely used for treatment of information in health and application tools of medical informatics. The objective of this paper is to present a MeSH-thesaurus conversion method to optimize complex operations in scientific researches MeSH-based. To exemplify we used an epistemology of medical informatics research, that seeks to understand if the medical informatics is characterized as science, technology or a mix of both, based on scientific literature, MeSH and of computational methods and tools criteria.

Key-words: Medical Informatics, Health Information, Thesaurus, Terminology, Text Mining.

Introdução

O estudo e uso de terminologias têm trazido importantes avanços para áreas interdisciplinares à informática em saúde, tais como inteligência artificial, mineração de dados, mineração de textos, busca e recuperação de informação, entre outras áreas da computação, amplamente utilizadas em aplicações para as Ciências.

Como tipos de terminologias, podemos citar os vocabulários controlados, os cabeçalhos de assuntos e os tesouros, os quais têm por objetivo a indexação, classificação, busca e recuperação de documentos, a partir de processos de análise e síntese. Vocabulários controlados são simples listas de palavras-chave com ordenação seqüencial ou alfabética, porém, sem nenhum tipo de relação e um controle mínimo destas palavras. Os cabeçalhos de assunto também são listas de termos, mas com maior controle sobre os termos e agregando relações diretas entre estes.

Os tesouros, bem mais sofisticados, apresentam avanço considerável em relação aos cabeçalhos de assunto, pois apresentam controle persistente e relações de vários tipos entre os termos. Por isso mesmo, os tesouros vêm ganhando cada vez mais espaço como instrumento de indexação e classificação de informação, em substituição aos vocabulários controlados e cabeçalhos de assunto.

O que diferencia um tesouro de um cabeçalho de assunto são os tipos de relações existentes entre os termos. Um cabeçalho de assunto apresenta somente relações hierárquicas entre os termos, enquanto no tesouro, além das relações hierárquicas, também são encontradas relações de equivalência e de associação. Com isto, a rede de relacionamentos entre os termos que não fazem parte de uma mesma hierarquia se tornar mais rica e sofisticada, o que vai refletir tanto nas estratégias de formulação da pesquisa quanto nos resultados da busca por informação a partir de um termo do tesouro.

Conceituação do Tesouro

Um tesouro pode ser definido como um vocabulário controlado que representa hierarquias, relações de equivalência, pertinência e associações entre os termos, com objetivo de auxiliar o usuário potencial a encontrar a informação de que necessita com a menor margem de erro possível. [1].

Os termos de um tesouro podem ser compostos por uma única palavra ou por várias palavras, formando um termo composto. Os termos de um tesouro são comumente denominados termos descritores, que Lancaster define como termos atribuídos por um indexador a um documento para descrever seu assunto [2].

As relações hierárquicas de um tesouro nada mais são que relações de ordenação entre os termos, isto é, envolvem a superordenação (acima de), subordinação (abaixo de) e coordenação (na mesma ordem, igual a). Fazendo uma analogia com uma família, pode-se dizer que uma mãe é superordenada em relação a seus filhos, enquanto um filho é subordinado à sua mãe e os filhos de uma mesma mãe, entre si, são coordenados, ou irmãos, conforme representado na Figura 1.

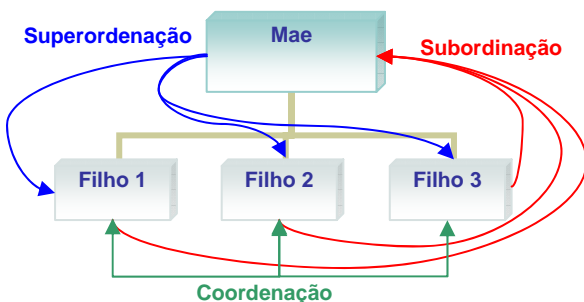


Figura 1. Relações hierárquicas de um tesouro.

As relações de equivalência envolvem o estudo e delimitação de termos diferentes com um mesmo significado e termos idênticos com significados diferentes, entre outras relações de equivalência entre termos já estabelecidas pela gramática das línguas, ou seja, sinônimos, antônimos, parônimos e homônimos. Junto às relações de equivalência são estabelecidas as relações de pertinência, que envolvem o estabelecimento de um termo padrão, com conceito e escopo bem definidos. Desta forma, fica instituído que o termo padrão será pertinente e seus sinônimos proibidos. Isto impede a pesquisa pelos sinônimos, mas sempre remete o usuário, ao utilizar um termo proibido, ao termo pertinente ou permitido. A Figura 2 apresenta uma representação dessas relações.

As relações associativas entre termos de um tesouro são aquelas que não se enquadram nas relações hierárquicas, nem nas de pertinência ou equivalência e ainda assim, permanecem e

são importantes para a recuperação da informação. Continuando a analogia de família, pensemos na seguinte situação: uma mulher se casa com um homem, sendo provenientes de famílias diferentes e têm filhos deste casamento. O casamento é uma espécie de relação associativa entre o homem e a mulher, que foge da hierarquia da família.

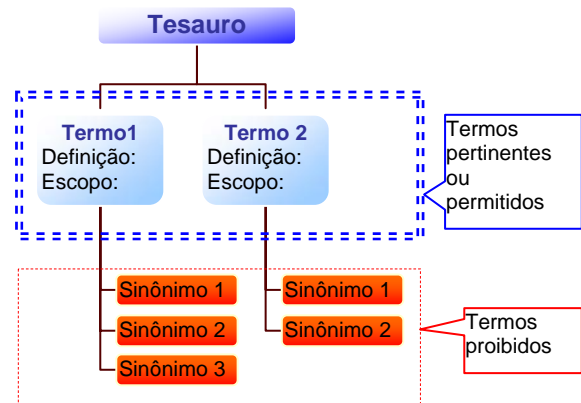


Figura 2. Relações de equivalência e pertinência de um tesouro.

Busca com Tesouro

Compreendido o conceito de um tesouro, podemos estabelecer a sua importância, que se fundamenta no potencial de auxílio ao usuário da informação em encontrar documentos de acordo com suas necessidades ou expectativas. Diferentes usuários podem expressar suas necessidades de informação, ainda que seja a mesma, usando uma linguagem diferente, por exemplo, sinônimos, abreviações, acrônimos etc [1].

O tesouro surge como uma alternativa para resolver estes problemas característicos do uso da linguagem natural, mapeando, por exemplo, os termos que representem o mesmo conceito, selecionando um termo apenas como padrão e os restantes como sinônimos, além de estabelecer relações entre estes termos e outros a estes relacionados.

O tesouro pode ainda representar a riqueza dos relacionamentos associativos e hierárquicos de tal maneira que usuários possam expressar sua necessidade de informação, limitando sua pesquisa a um nível de especificidade mais restrito ou mais amplo do que aquele usado pelo indexador, melhorando os resultados da busca [1].

Além disso, técnicas e ferramentas da mineração de textos em estudo na inteligência artificial e na lingüística computacional vêm utilizando tesouros como instrumentos para extração automática de informação, a partir de conjuntos de textos (corpus) que resultam numa série de aplicações, tais como a indexação automática, a tradução automática interlíngua, a sumarização de textos etc [1]. Com isto, pode-se compreender

o motivo pelo qual o tesouro vem sendo tão valorizado e utilizado em detrimento de terminologias mais simplificadas como os cabeçalhos de assuntos.

MeSH

Concentrando nossa abordagem em aplicações de informática em saúde, deparamo-nos com o Medical Subject Headings (MeSH) [3], que é também um destes instrumentos terminológicos largamente utilizados e cujo domínio de estudo e atuação está delimitado às Ciências da Saúde. O MeSH é um cabeçalho de assunto especializado em ciências da saúde, desenvolvido, publicado e disponível online na internet pela National Library of Medicine (NLM). É atualizado dinamicamente por especialistas de várias áreas do conhecimento.

No MeSH, um descritor representa uma classe de conceitos, enquanto um conceito representa uma classe de termos sinônimos. A organização do MeSH se dá em 16 categorias de assuntos, sendo que cada uma se divide em subcategorias, nas quais os descritores subordinados são organizados hierarquicamente numa relação do mais genérico para o mais específico [4].

Os principais usos do MeSH são a indexação de artigos, a classificação de itens de informação e a pesquisa em bancos de dados de literatura científica em saúde, que tenham sido indexados pelo MeSH.

A terminologia MeSH oferece um modo consistente para recuperar informação permitindo o uso de diferentes terminologias para os mesmos conceitos. A organização dos termos descritores é feita em uma estrutura hierárquica, a qual oferece um modo efetivo para se encontrar palavras-chave apropriadas para uma pesquisa. O seu idioma principal é o inglês.

Para mostrar a importância do MeSH podemos citar a base de dados de literatura científica em saúde MEDLINE/PubMed [5], indexada pelo MeSH, que contém mais de 16 milhões de registros indexados com taxa de crescimento de 500.000 artigos/ano, cobrindo aproximadamente 4.600 revistas biomédicas internacionais.

O MeSH, em sua versão completa, está disponível para download nos formatos XML e TXT para propósitos específicos de pesquisa científica e para uso por centros de informação para indexação de seus itens informacionais. Também encontra-se disponível online para pesquisa de termos e suas relações [6].

Objetivos

Este trabalho apresenta uma proposta de modelagem de dados para o cabeçalho de assunto MeSH para transformá-lo em um tesouro, utilizando como repositório um banco de dados relacional e técnicas de modelagem e projeto de banco de dados. Apresenta ainda os benefícios do uso do MeSH em forma de tesouro que podem ser maiores que seu uso em forma de cabeçalho de assunto.

O objetivo desta conversão é a otimização de operações complexas efetuadas com os conjuntos de termos MeSH. Como exemplo, será considerado o uso do MeSH em sua versão completa como instrumento para pesquisa científica de epistemologia da informática em saúde. Tal pesquisa visa identificar em um vasto conjunto de artigos científicos os termos descritores ou palavras-chave pertinentes à área de informática em saúde, a partir dos termos MeSH. O conjunto resultante de descritores deverá ser categorizado como termo científico ou termo tecnológico. Então será possível inferir se a informática em saúde se caracteriza mais como ciência ou como tecnologia ou uma mescla de ambos.

Metodologia

Fizemos o download do MeSH da internet em versão TXT e o transformamos para uma versão de banco de dados relacional, mantendo o modelo relacional estabelecido pela NLM na sua origem. Entretanto, a complexidade do modelo relacional da NLM com um grande número de tabelas pode ser um entrave às complexas operações que se pretende realizar no processamento de textos, como se pode observar na Figura 3

Devido a isto, foi desenhado um modelo relacional de banco de dados, apresentado na Figura 4, que atenda às principais características de um tesouro de forma simplificada e buscando seguir as 4 primeiras formas normais, com base em um projeto conceitual e lógico adequados.

Um modelo simplificado de banco de dados relacional para armazenamento e manipulação de um tesouro poderá permitir melhor processamento das operações, especialmente consultas SQL, necessárias ao estudo que se pretende aqui: compreender se a informática em saúde é uma ciência ou uma tecnologia ou uma ou uma mescla de ambos.

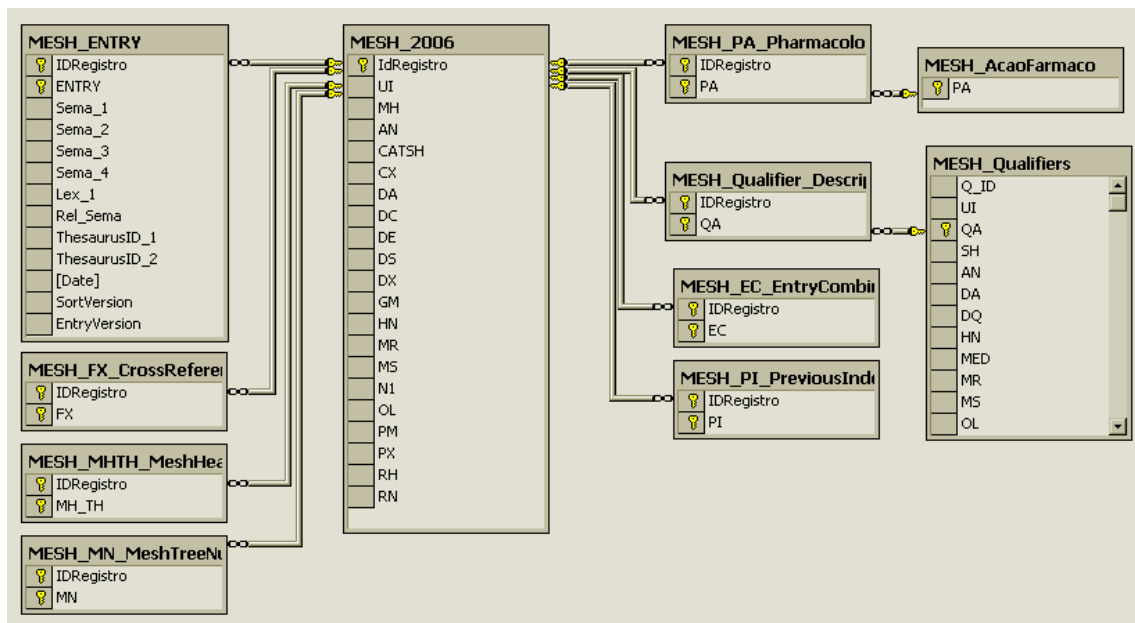


Figura 3. Diagrama entidade-relacionamento do MeSH em formato original.

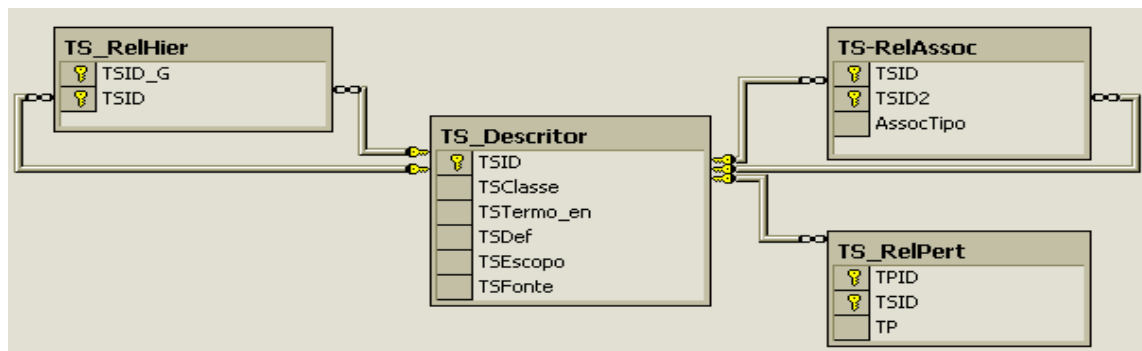


Figura 4. Diagrama Entidade-Relacionamento de um tesauro genérico.

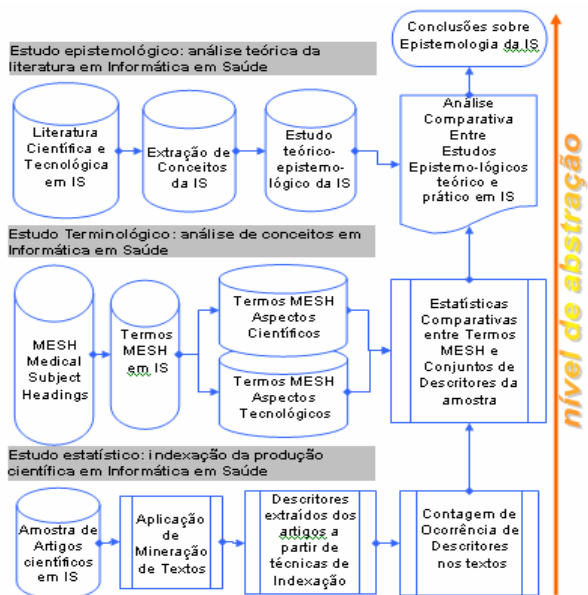


Figura 5. Processo epistemológico para delimitação científica ou tecnológica da informática em saúde.

O processo de caracterização da informática em saúde como ciência ou tecnologia se divide em etapas, representadas na Figura 5:

- Um estudo estatístico que visa analisar uma amostra de artigos científicos em informática em saúde utilizando técnicas de mineração de textos e indexação e a contagem de um subconjunto de descritores resultantes;
- A seguir, o estudo terminológico utilizando o MeSH vai permitir a seleção e análise de conceitos em informática em saúde e a comparação entre estes termos e aqueles resultando do estudo estatístico;
- Um estudo epistemológico envolvendo a análise e extração de conceitos da literatura científica em informática em saúde por meio de leituras e processos de análise e síntese, que vai permitir a delimitação dos conceitos.

Uma análise comparativa entre os estudos estatístico-terminológico, de cunho prático, e o estudo epistemológico, de cunho teórico, permitirá chegar a alguma conclusão sobre a epistemologia da informática em saúde. Compreende-se que um modelo genérico de banco de dados relacional para tesauro pode contribuir para diminuir a quantidade de objetos e relacionamentos, simplificando o banco de dados para consultas complexas que envolvem JUNÇÃO (JOIN). O mesmo

pode ser feito com o objeto MeSH_2006, que contém termos descritores MeSH, com grandes quantidades de textos completos de artigos científicos para rastreamento e recuperação destes termos dentro do texto completo, no escopo da informática em saúde. O resultado dessas contagens possibilita a análise quantitativa e, posteriormente, qualitativa, do uso efetivo de termos da informática em saúde, sejam eles de cunho científico ou tecnológico.

Resultados

O resultado desta pesquisa envolve a criação de um projeto conceitual de tesauro modelado em forma de banco de dados relacional e da análise comparativa entre consultas realizadas no MeSH em seu formato original e no novo modelo proposto.

O projeto conceitual proposto pode ser demonstrado no diagrama entidade-relacionamento representado na Figura 6.

O que se pode observar no diagrama proposto é que muitos dos objetos presentes no modelo tradicional do MeSH, tais como MESH_Entry, MESH_PA, MESH_Qualifier, MESH_EntryCombination e outros, tiveram seus conjuntos de dados introduzidos no objeto TS_Descriptor, resultando numa tabela única de termos descritores, cujas relações se estabelecem de forma recursiva por meio das seguintes tabelas relacionais:

- TS_RelHier, que representa as relações hierárquicas entre os descritores (descritor é subordinado a);
- TS_RelAssoc, que representa as relações associativas entre os descritores (descritor é associado a);
- TS_RelPert, que representa as relações de equivalência e de pertinência entre os descritores (descritor é usado para).

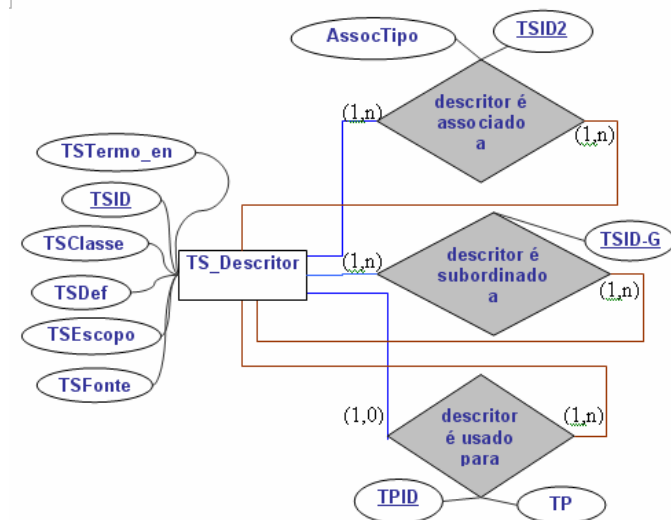


Figura 6. Proposta de diagrama entidade-relacionamento para tesauro MeSH.

Com isto, temos apenas 4 objetos no banco de dados, sendo TS_Descriptor o objeto principal com o qual trabalharemos a maior parte do tempo para elaboração de consultas. Os outros 3 objetos são relacionamentos recursivos entre os descritores, ou seja: suas relações hierárquicas, associativas e de equivalência/pertinência. A seguir, um exemplo de como se dão as relações entre os termos.

Os termos Informática em Saúde e Telemedicina são descritores cadastrados na tabela TS_Descriptor. Porém, existe entre estes 2 termos uma relação hierárquica, ou seja, o termo “Informática em Saúde” é superordenado em relação ao termo “Telemedicina” e, por sua vez, o termo “Telemedicina” é subordinado ao termo “Informática em Saúde”. Para que estas relações sejam estabelecidas, não se faz necessário recadastrar os mesmos termos na tabela TS_RelHier, pois esta tabela apresenta 2 chaves estrangeiras que formam uma chave primária composta. Estas chaves estrangeiras que compõem a tabela TS_RelHier são provenientes do atributo chave TS_ID da tabela TS_Descriptor:

- TSID_G, que representa o termo superordenado, que é uma chave estrangeira proveniente do atributo TSID da tabela TS_Descriptor
- TSID, que representa o termo subordinado, que é uma chave estrangeira também proveniente do atributo TSID da tabela TS_Descriptor.

Supondo que o TSID do termo “Informática em Saúde” é 25 e o TSID do termo “Telemedicina” é 32, temos o seguinte registro na tabela TS_RelHier:

TSID_G	TSID
25 (Informática em Saúde)	32 (Telemedicina)

O mesmo ocorrerá na tabela relacional TS_RelAssoc, que só conterá os devidos relacionamentos entre termos descritores não-hierárquicos. Com a tabela TS_RelPert, haverá apenas as relações entre termos descritores (TSID) e seus equivalentes (TP – termo proibido), tal como no exemplo:

TSID (ID do termo permitido)	TP (termo equivalente mas proibido)
25 (Informática em Saúde)	eSaúde
25 (Informática em Saúde)	informática médica

Desta forma, foram reduzidas as quantidades de objetos do sistema relacional e, consequentemente, das operações com estes objetos e seus conjuntos de dados. Mais que isso, a maioria das operações a serem realizadas com o tesauro MeSH serão feitas apenas com o objeto TS_Descriptor. Porém, mesmo quando forem usadas as tabelas relacionais TS_RelAssoc, TS_RelHier e TS_RelPert, serão utilizadas as suas respectivas chaves estrangeiras, que são

numéricas, para cálculos e contagens. Como o computador tem mais facilidade e agilidade para trabalhar com os números, neste caso os IDs que são os códigos identificadores de registros de bancos de dados relacionais, as consultas e operações a serem realizadas também com as tabelas relacionais serão mais eficientes.

Discussão e Conclusões

Neste artigo ressaltamos a importância do estudo e uso das terminologias para diversas áreas de pesquisa que lidam com a informação e documentação, inclusive a informática em saúde. Elencamos tipos de terminologias utilizados como instrumento para indexação, classificação, busca e recuperação de informação, tais como os vocabulários controlados, os cabeçalhos de assunto e os tesouros, que ressaltamos devido à sua importância e utilidade em aplicações diversas.

O tesouro é uma terminologia sofisticada que apresenta um conjunto de termos de um domínio específico assim como as relações hierárquicas, associativas e de equivalência e pertinência entre estes termos, formando uma rede de informação ao mesmo tempo íntegra e flexível para busca de informação. Por isto, acreditamos que o tesouro seja o instrumento terminológico mais adequado à pesquisa científica que envolva a indexação, busca e recuperação de informação.

Vimos que o MeSH é um importante instrumento para indexação, classificação, busca e recuperação de informação para as ciências da saúde, além de ser amplamente utilizado em ferramentas de aplicação da informática em saúde. Sendo o MeSH, em sua forma original, considerado como um cabeçalho de assunto devido à sua estrutura interna, nosso objetivo foi apresentar um método de transformação do MeSH em um tesouro, no intuito de otimizar operações complexas efetuadas com os conjuntos de termos MeSH em pesquisas científicas.

Como exemplo, foi utilizada uma pesquisa sobre a epistemologia da informática em saúde, cujo objetivo é identificar em um conjunto de artigos científicos os termos pertinentes à área de informática em saúde, a partir dos termos MeSH, sendo cada um dos termos resultantes categorizado como científico ou tecnológico, de onde se poderá inferir se a informática em saúde se caracteriza mais como ciência, tecnologia ou uma mistura de ambos.

Como resultado, apresentamos um diagrama entidade-relacionamento contendo a estrutura de dados do MeSH transformado em modelo de tesouro. Este modelo visa contribuir na redução da quantidade de objetos e relacionamentos do sistema MeSH, e na simplificação do banco de dados para execução de consultas complexas, especialmente as que envolvem JUNÇÃO.

Com isto, esperamos contribuir com novas pesquisas em informática em saúde que utilizem como instrumento a terminologia MeSH, porém, incorporando a esta terminologia os conceitos e a sofisticação de um tesouro, que nos parecem robustos e flexíveis, para indexação, classificação, busca e recuperação de informação.

Agradecimentos

Os autores agradecem ao Prof. Dr. Mauro Biajz, professor do Departamento de Computação da Universidade Federal de São Carlos (UFS-Car), que motivou o desenvolvimento deste trabalho.

Referências

- [1] Ebecken, N.F., Lopes, M.C.S., Costa, M.C.A. (2003), "Mineração de textos", In: *Sistemas inteligentes*, Org.: Solange de Oliveira Rezende, Barueri, SP: Manole, p. 337-370.
- [2] Lancaster, F.W. (1972), "Vocabulary control for information retrieval", Washington, Information Resources Press.
- [3] National Library of Medicine (2005), *MeSH: Medical Subject Headings*, USA, jan. [\[http://www.nlm.nih.gov/mesh/meshhome.html\]](http://www.nlm.nih.gov/mesh/meshhome.html). 2 Junho 2006.
- [4] National Library of Medicine (2005), *MeSH Tree Structures*, USA, nov. [\[http://www.nlm.nih.gov/mesh/intro_trees2006.html\]](http://www.nlm.nih.gov/mesh/intro_trees2006.html). 2 Junho 2006.
- [5] NCI (2006), *PubMed*, EUA. [\[http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed\]](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed). 2 Junho 2006.
- [6] NLM (2005), *Medical Subject Headings: files available to download*. Nov. [\[http://www.nlm.nih.gov/mesh/filelist.html\]](http://www.nlm.nih.gov/mesh/filelist.html). 2 Junho 2006.

Contato

Eliane Colepícolo, Prof. Dr. Ivan Torres Pisa e os demais autores recebem correspondências no endereço: Departamento de Informática em Saúde, UNIFESP, Rua Botucatu, 862, CEP 04023-062, Vila Clementino, São Paulo, SP. Telefones (11) 5576-4521 e 5574-5234. Os e-mails são colepicolo-pg@dis.epm.br e ivapisa@dis.epm.br.