Reengenharia de Tesauro: caso do Thesagro

Benildes C. M. S. Maculan
Universidade Federal de Minas Gerais
benildes@gmail.com

Gercina A. B. O. Lima
Universidade Federal de Minas Gerais
limagercina@gmail.com

Ivo Pierozzi Jr.
Embrapa Informática Agropecuária
ivo.pierozzi@embrapa.br

Leandro H. M. Oliveira
Embrapa Informática Agropecuária
ivo.pierozzi@embrapa.br

1 Introdução

Os instrumentos de representação da informação para uso em ambiente digital têm sido agrupados sob a designação de Sistemas de Organização do Conhecimento (SOC). Nesse ambiente, os SOCs “são vistos como esquemas que visam organizar, gerenciar e recuperar informações” (VICKERY, 2007, *on-line*). Dessa forma, eles têm uma única finalidade: “organizar conteúdos para apoiar a recuperação de itens relevantes, disponibilizados na base de dados de uma biblioteca digital” (HODGE, 2000, p. 9). Para cumprir essa finalidade, é importante imprimir maior riqueza semântica na estrutura conceitual dos SOCs.

Tradicionalmente, o tesauro é um tipo de SOC que possui uma estrutura conceitual semântica, pois é composto por um conjunto de conceitos que estão ligados entre si por diferentes tipos de relações. Esses relacionamentos formam uma rede semântica de relações entre termos e conceitos, representados por relações de equivalência (termos sinônimos e variações linguísticas), identificados pelos símbolos USE e UP (usado para), hierárquicas (conceitos ordenados e agrupados por níveis diferentes de generalidade e especificidade), identificados pelos símbolos BT (termo genérico; *broader term*) e NT (termo específico; *narrower term*), e associativas (outras relações que não de equivalência ou hierárquicas), identificadas pelo símbolo RT (termo relacionado; *related term*). Os estudos sobre os princípios aplicados na construção de tesauros vêm evoluindo (MOTTA, 1987; CAMPOS, 1995; CAMPOS; GOMES, 2003, entre outros), e já há iniciativas que propõem a especificação dos diferentes tipos de relações em tesauros.

Entre essas iniciativas, prevalecem as investigações sobre modelos de conversão de tesauros em ontologias de domínio, visando ao reuso do conhecimento já estruturado dos tesauros. Nesse contexto, este artigo relata parte dos resultados da aplicação do modelo desenvolvido por Dagobert Soergel *et al.* (2004) e Lauser *et al.* (2006) na conversão do *Thesaurus* Agrícola Nacional (Thesagro) em um instrumento mais formalizado, de tal forma que os relacionamentos de sua estrutura conceitual estejam explicitados para o usuário. Esse tesauro foi escolhido por ser o único tesauro brasileiro da área da Agropecuária, uma vez que este estudo foi realizado com o apoio da Embrapa Informática Agropecuária (Embrapa), unidade sediada em Campinas-SP, como parte de uma parceria firmada entre essa instituição, a Universidade Federal de Minas Gerais (UFMG) e o Grupo de Pesquisa Protótipo Mapa Hipertextual (MHTX).

2 O modelo de conversão de tesauro

A proposta do modelo utilizado na conversão do tesauro Thesagro tem seus procedimentos detalhados em dois artigos: (1) Soergel *et al.* (2004) e (2) Lauser *et al.* (2006). Esse modelo foi demonstrado pelos autores na conversão de uma amostra do tesauro Agrovoc e teve como objetivo a criação de uma ontologia pesada de domínio, composta por classes, atributos e relacionamentos entre entidades, todos expressos em *Web Ontology Language, Description Logic* (OWL DL). Como a proposta dos autores foi

apresentada de forma gradual, foi possível separar a fase conceitual de modelagem do tesauro da fase de automatização do processo de conversão do tesauro em uma ontologia de domínio. Portanto, os resultados apresentados neste artigo relatam os procedimentos intelectuais de refinamento e explicitação das relações existentes na estrutura conceitual do tesauro Thesagro.

A principal característica desse modelo de reengenharia de tesauros é possibilitar a individualização da modelagem em cinco níveis de entidades: conceito, termo ou lexicalização, *string* ou variantes, notas de escopo e relacionamentos. Dessa maneira, cada nível indica diferentes tipos de informação e os relacionamentos, considerados a espinha dorsal do tesauro, podem ser atribuídos entre entidades de mesmo nível (por exemplo, entre diferentes conceitos) ou entre entidades de níveis distintos (por exemplo, entre termos e *strings*), conforme a seguir:

conceito para conceito	é_uma (hierarquia); praga_de
termo para termo	é_sinônimo_de; é_tradução_de
conceitos para termos	tem_lexicalização (liga os conceitos a seu representante lexical)
termo para *string*	tem_acrônimo; tem_variação_ortográfica; tem_abreviatura (liga os termos com suas formas variantes)

O modelo é composto por três etapas básicas: (1) definição da estrutura do tesauro convertido, utilizando um tesauro existente no domínio a ser trabalhado; (2) coleta de terminologia e outras informações, a partir de um ou mais tesauros, no domínio a ser modelado; (3) edição do tesauro, com a reformulação do tesauro existente, transformando a sua estrutura em uma rede conceitual mais semântica.

2.1 Aplicação das etapas do modelo de conversão de tesauro

Para a aplicação do modelo de conversão de tesauro, foi necessário fazer um planejamento inicial no qual foi constituída uma equipe de trabalho composta por cinco integrantes: um modelador (profissional da informação), um especialista do domínio, dois terminólogos e um tecnólogo.

A primeira etapa, da definição da estrutura do tesauro convertido, abrangeu o mapeamento das características do Thesagro. Verificamos que esse tesauro possui cerca de 9.400 termos descritores, preferidos e não-preferidos, e ele é considerado um

tesauro de alta especificidade e de amplo escopo. De forma semelhante a outros tesauros tradicionais, o Thesagro apresenta os três relacionamentos básicos, utilizando símbolos na língua inglesa:

Relações de equivalência:	USE e USED FOR (UF)
Relações hierárquicas:	BROADER TERM (BT) e NARROWER TERM (NT)
Relações associativas:	RELATED TERM (RT)

Pela análise do tesauro Thesagro, observamos a presença de cerca de 2.000 descritores preferidos não pertencentes a qualquer *cluster* na estrutura hierárquica do Thesagro, sendo considerados descritores órfãos. Essa falta de vinculação entre conceitos pode, algumas vezes, tornar a semântica da estrutura do tesauro menos compreensível para o usuário.

Também na primeira etapa incluiu a determinação da subárea da Intensificação Agropecuária como o recorte temático para a modelagem. A conceitualização desse tema segue da teoria de Boserup (1965) segundo a qual existe uma relação entre a dinâmica da população de uma região, com seu crescimento ou não, o meio ambiente e a utilização de tecnologia na produção agropecuária, originando o aumento ou a manutenção de uma mesma em um dado tempo e lugar. A modelagem dessa subárea teve por base inicial uma taxonomia elaborada pelos especialistas da Embrapa, composta por cerca de 600 conceitos e que foi estruturada em nove categorias: (1) agricultura extensiva; (2) agricultura intensiva; (3) material e métodos; (4) ambiente; (5) agronomia; (6) território e paisagem; (7) socioeconomia; (8) espaço e tempo; (9) instituições. Entre os conceitos da taxonomia foi definida uma amostra composta por 30 conceitos representativos de cada uma das nove categorias e, assim, da subárea da Intensificação Agropecuária como um todo.

Na segunda etapa, da coleta de terminologia, foram utilizados como insumos terminológicos a taxonomia e três tesauros com escopo da agropecuária: o Thesagro, o Agrovoc e o *National Agricultural Library* (NAL). Para essa atividade foi realizada uma comparação entre os conceitos da amostra e a terminologia existente na taxonomia e em cada um dos três tesauros selecionados. Isso foi feito a partir de duas listas: (1) lista composta pelos 30 conceitos da amostra, em português brasileiro, adicionando-se as expressões desses termos no singular e plural; (2) lista composta pelos 30 conceitos da

amostra, traduzidos para o inglês, adicionando-se as expressões dos termos no singular e plural, e também na sua forma inversa (adjetivo + substantivo), por essa inversão ser comum na língua inglesa. A comparação terminológica foi efetivada de forma semiautomática, a partir de uma análise intelectual dos instrumentos e de um processamento automático de comparação. Esse processamento foi realizado utilizando o Extrator de Termos e Estruturas Conceituais Agrícolas Multilíngue (ETECAM), que permitiu resgatar os termos coincidentes com as duas listas nos instrumentos, assim como seus *clusters* semânticos, separadamente.

A terceira etapa envolveu a edição da remodelagem da estrutura conceitual do Thesagro, utilizando o *software* Termos Eletrônicos (e-Termos), que é uma ferramenta que dá suporte à criação e gestão de produtos terminológicos para distintos fins (ensino, glossários, vocabulários controlados). O sistema e-Termos foi desenvolvido como um ambiente computacional colaborativo *web*, que pode ser utilizado de forma gratuita, mas restrita a usuários cadastrados. Ele é composto por seis etapas, com procedimentos automatizados e semiautomatizados, tendo por base os princípios da Teoria Comunicativa da Terminologia (TCT), desenvolvida por Cabré (1999). A TCT possui um construto teórico que defende uma semelhança entre os sistemas da linguagem de especialidade e das línguas gerais, pois ambas são regidas pelas mesmas regras e são caracterizadas pelos mesmos fenômenos de sinonímia e variação linguísticas. Esses princípios minimizam a rigidez normativa advinda da Teoria Geral da Terminologia (TGT), desenvolvida por Wüster (1998), cujos fundamentos são geralmente utilizados na construção de tesauros.

O sistema e-Termos possibilitou a criação de uma Base Definicional que compilou e armazenou excertos definitórios explicativos e/ou sobre os conceitos da amostra. O conteúdo desses excertos auxiliou a confecção de um glossário com as definições dos conceitos da amostra. Além disso, o e-Termos também permitiu a elaboração de fichas terminológicas para esses conceitos, que foi composta por 38 campos semânticos, entre os quais destacamos os campos para as definições dos conceitos (do especialista, modelador e final), para informações enciclopédicas e de glosa, para notas de escopo e para os termos em relação de equivalência, de variação linguística, assim como campos para os conceitos em relação hierárquica e associativa.

Na construção do sistema de conceitos para os conceitos da amostra e seus *clusters* semânticos foram utilizados 44 diferentes tipos de relações. Isso gerou um desdobramento que totalizou a representação de cerca de 600 relacionamentos: de gênero e suas espécies, do todo e suas partes, de equivalências, de strings (variações) e associativas.

Para uma exemplificação da estrutura atual do Thesagro e a propostas de reformulação utilizando o modelo de conversão de tesauros, a seguir apresenta-se a atual modelagem do descritor GATO no Thesagro:

GATO *BT* MAMÍFERO DOMÉSTICO *NT* GATO ANGORÁ *NT* GATO DO MATO *RT* FELIS CATTUS DOMESTICUS *RT* FELIS DOMESTICA	FELIS CATTUS DOMESTICUS *RT* GATO FELIS DOMESTICA *RT* GATO

Na estrutura atual do Thesagro nota-se, por exemplo, que o descritor MAMÍFERO DOMÉSTICO é um conceito mais genérico e que os descritores GATO ANGORÁ e GATO DO MATO são conceitos mais específicos de GATO. No que diz respeito aos outros relacionamentos, é possível perceber que o descritor preferido GATO está em uma relação associativa com os descritores preferidos FELIS CATTUS DOMESTICUS e FELIS DOMESTICA, sem que esteja claro o tipo específico de relação associativa.

Aplicando o modelo de conversão de tesauros, a reformulação da modelagem do descritor GATO ficou com a seguinte configuração:

GATO *temTermoGenérico* MAMÍFERO DOMÉSTICO *temTermoEspecífico* GATO ANGORÁ *temTermoEspecífico* GATO DO MATO *temNomeCientífico* FELIS CATTUS DOMESTICUS *temNomeCientífico* FELIS DOMESTICA	FELIS CATTUS DOMESTICUS *temNomePopular* GATO FELIS DOMESTICA *temNomePopular* GATO

Com essa reformulação, o tipo de relação que foi estabelecida entre os conceitos ficou explícito, facilitando o entendimento da estrutura semântica. Assim, qualquer usuário pode apreender que o descritor MAMÍFERO DOMÉSTICO é um termo mais genérico e está em uma relação de gênero-espécie com GATO que, por sua vez, está em uma relação também de gênero-espécie com GATO ANGORÁ e GATO DO MATO.

Ademais, agora percebemos claramente que os descritores FELIS CATTUS DOMESTICUS e FELIS DOMESTICA são nomenclaturas científicas do descritor GATO, que é um nome popular deste animal. Dessa forma, podemos afirmar que a modelagem atual do Thesagro deixa perdida a função semântica que conecta o descritor GATO com os descritores que representam as nomenclaturas científicas. Esse tipo de estruturação se repete na modelagem de todos os organismos vivos incluídos na terminologia do atual Thesagro e explica a existência dos cerca de 2.000 descritores considerados órfãos. Com essa forma de modelagem, a representação atual do Thesagro mantém disperso o sentido semântico da sua estrutura conceitual, dificultando a compreensão do conhecimento do domínio, a não ser quando o usuário já possui esses conhecimentos *a priori*

Com a reformulação dos conceitos da amostra e de seus *clusters* semânticos foram representadas 286 ocorrências de relações hierárquicas, sendo 225 relações de gênero-espécie (com 52 termos gerais e 173 termos específicos) e 61 relações todo-partes (com 22 termos gerais e 39 termos específicos), o que mostra que a nova estrutura do Thesagro mantém a sua característica de especificidade temática. Além disso, foram representados 232 relacionamentos associativos, confirmando que a subárea da Intensificação Agropecuária tem natureza complexa, com muitos conceitos inter-relacionados.

Os resultados evidenciaram que a expressão explícita das relações entre os pares de entidades (conceitos, termos, *strings* e notas de escopo) refinou a semântica da estrutura do tesauro convertido, dando subsídios para facilitar a interoperabilidade entre diferentes tesauros ou sistemas. Ressaltamos que os resultados completos pela pesquisa podem ser consultados na tese da autora principal deste artigo.

3 Considerações finais

O desenvolvimento deste estudo teve como ponto de partida o enriquecimento semântico da estrutura conceitual de um tesauro tradicional, visando a torná-lo um tipo de SOC, denominação que abrange os instrumentos que podem ser compreendidos pela máquina.

Durante o percurso de pesquisa, ficou evidente que o trabalho de construção ou de reformulação de um tesauro é bastante complexo, o que exige o envolvimento de distintos profissionais, tais como bibliotecários, terminólogos e especialistas. Além disso, verificamos a importância do atendimento às necessidades do usuário, uma vez que a terminologia de um tesauro, ainda que seja uma linguagem de especialidade, deve priorizar a linguagem de busca do usuário.

Percebemos, também, que o desenvolvimento da estrutura semântica de um tesauro deve ser orientado por normas e padrões internacionais para construção de tesauros, pois eles estabelecem princípios para uma representação mais formal, o que pode auxiliar na interoperabilidade entre diferentes vocabulários e sistemas.

Consideramos como principal contribuição desta pesquisa a validação de um modelo de conversão de tesauros tradicionais, que pode ser aplicado como solução de integração de dados em diferentes sistemas. Avaliamos que isso pode impactar positivamente na área da Organização da informação, fazendo avançar os estudos acerca de metodologias que podem ser utilizadas na construção de tesauros.

Referências

BOSERUP, E. **The conditions of agricultural growth:** the economics of agrarian change under population pressure. Chicago: Aldine, 1965.

CABRÉ, M. T. **La terminología:** representación y comunicación. Barcelona: Institut Universitari de Linguística Aplicada, Universitat Pompeu Fabra, 1999.

CAMPOS, M. L. A. Linguagens documentárias: núcleo básico de conhecimento para seu estudo. **R. Esc. Biblioteconomia UFMG**, Belo Horizonte, v. 24, n. 1, p. 52-62, jan./jun. 1995.
CAMPOS, M. L. A.; GOMES, H. E. Organização de domínios de conhecimento e os princípios ranganathianos. **Perspect. Ci. Inf.**, Belo Horizonte, v. 8, n. 2, jul./dez. 2003.

DAHLBERG, I. Teoria do conceito. Tradução Astério Tavares Campos. **Ci. Inf.,** Rio de Janeiro, v. 7, n. 2, p. 101-107, 1978.

LAUSER, B. et al. From AGROVOC to the Agricultural Ontology Service: Concept Server an OWL model for creating ontologies in the agricultural domain. In: INTERNATIONAL CONFERENCE ON DUBLIN CORE AND METADATA APPLICATIONS, 2006, Colima, Mexico. **Proceedings…** México: DCMI, 2006.

MOTTA, D. F. da. **Método relacional como nova abordagem para a construção de tesauros.** 1987. 89f. Dissertação (Mestrado em Ciência da Informação) – Instituto Brasileiro de Informação em Ciência e Tecnologia, Rio de Janeiro, 1987.

SOERGEL, D. et al. Reengineering thesauri for new applications: the AGROVOC example. **Journal of Digital Information**, v. 4, n. 4, 2004.

WÜSTER, E. **Introducción a la teoría general de la terminología y a la lexicografía terminológica.** Barcelona: IULA, 1998.