Aplicação de reengenharia de tesauro: modelagem do THESAGRO

Reengineering thesaurus application: modeling THESAGRO

Benildes C. M. S. Maculan (1), Gercina A. B. O. Lima (2), Ivo Pierozzi Jr. (3), Leandro H. M. Oliveira (4)

(1) Escola de Ciência da Informação, Universidade Federal de Minas Gerais (ECI/UFMG), Av. Presidente Antônio Carlos, número 6627, Gabinete 4030, Pampulha, Belo Horizonte, MG, Brasil, CEP 31270-901, benildes@gmail.com (2) limagercina@gmail.com
(3) Embrapa Informática Agropecuária, Av. André Tosello, 209, Cidade Universitária, Campinas, SP, CEP 13083-886, ivo.pierozzi@embrapa.br. (4) leandro.oliveira@embrapa.br

Resumo

Este estudo investigou e aplicou um modelo de reengenharia de tesauros tradicionais para tornar o sistema de relações entre os conceitos em uma estrutura semântica rica. O modelo selecionado para a aplicação foi desenvolvido por Soergel *et al.* (2004) e Lauser *et al.* (2006). Esse modelo é composto por três etapas e envolve a melhoria e explicitação formal das relações semânticas em um tesauro. A reengenharia foi aplicada no tesauro brasileiro THESAGRO, do domínio da Agricultura, no recorte temático da Intensificação Agropecuária. A metodologia incluiu o uso das ferramentas: (a) Sistema e-Termos, para o gerenciamento da terminologia; e (b) Extrator de Termos, *software* que permite a comparação automática da terminologia de diferentes tesauros. Os resultados demonstraram a viabilidade da utilização do modelo aplicado na reengenharia de tesauros tradicionais, uma vez que permitiu modelar e obter uma estrutura mais semanticamente enriquecida. Concluiu-se que o refinamento das relações ajudou a organização do conhecimento da subárea temática modelada, o que pode facilitar a sua exploração pelo usuário final, assim como ser um importante elemento para a promoção da interoperabilidade entre diferentes tesauros.

Palavras-chave: Reengenharia tesauros. Sistema de organização do conhecimento. Thesauros. Modelagem conceitual. Modelo de conversão de tesauros.

Abstract

Research that aimed to study and apply a model of reengineering thesauri making it a conceptual structure semantically enriched. The method included the selection of the reengineering model developed by Soergel et al. (2004) and Lauser et al. (2006). This model consists of three stages and consists of the improvement and formal explanation of the semantic relationships in the conceptual structure of a traditional thesaurus. The reengineering was applied to the Brazilian thesaurus Thesagro from the Agriculture domain more specifically it was applied to the thematic frame of the agriculture intensification. In methodological course some tools were used: (a) e-Terms system: for the terminology management; and (b) Terms Extractor, a software that allows automatic comparison of the terminology of different thesauri. The results showed the viability of the use of the analyzed model in the reengineering of tradition thesauri because it allowed the imprinting of more semantics to the structure of the modeled thesaurus specifying the kind of relationship existing between concepts and terms. It was concluded that the possibility of refining the relations between concepts helped in the organization of the modeled domain and it can facilitate the exploration by the final users.

Keywords: Reengineering of thesaurus. Knowledge Organization System. Thesaurus. Conceptual Modeling. Thesaurus model conversion.

1. Introdução

No campo da Biblioteconomia e Ciência da Informação (BCI), os distintos tipos de instrumentos de representação, tais como taxonomias, listas de cabeçalho de assunto, tesauros, redes semânticas e ontologias, têm sido agrupados sob a designação de Sistemas de Organização do Conhecimento (SOC). Esse termo foi cunhado em 1998, durante a primeira Conferência da ACM *Digital Libraries*, em Pittsburgh (Pennsylvania), quando o *Networked Knowledge Organization Systems Working Group* propôs o uso do termo "*Knowledge Organization System*" (KOS). Para Vickery (2007, *on-line*) os SOCs "são vistos como esquemas que visam organizar, gerenciar e recuperar informações", para aplicações em ambientes digitais. Carlan (2010) afirma que os SOCs representam uma "denominação nova para as linguagens documentárias que

Maculan, B. C. M. S.; Lima, G. A. B. O.; Pierozzi Jr, I.; Oliveira, L. H. M. *Aplicação de reengenharia de tesauro: modelagem do Thesagro.* En XII Congreso ISKO España y II Congreso ISKO España-Portugal, 19-20 de noviembre, 2015, Organización del conocimiento para sistemas de información abiertos. Murcia: Universidad de Murcia.

agregam elementos incorporados nas inovações tecnológicas da era digital" (CARLAN, 2010, p. 29-30).

Hodge (2000) acrescenta que os SOCs são utilizados para "organizar conteúdos para apoiar a recuperação de itens relevantes, disponibilizados na base de dados de uma biblioteca digital" (HODGE, 2000, p. 9). Sendo assim, os SOCs são instrumentos, já tradicionais na área da Biblioteconomia, que podem ser utilizados para a representação e na recuperação de informações junto a aplicações tecnológicas em ambiente digital.

Segundo Soergel (1999), um SOC tem a função de ser um dicionário mono, bi ou multilíngue, para uso humano ou como base de conhecimento em uma aplicação em ambiente digital, para ser compreendido pela máquina. O autor afirma que os SOCs têm como objetivos:

(1) mapear domínios individuais, sendo um mapa semântico capaz de indicar os relacionamentos entre conceitos no domínio mapeado e servindo como uma ferramenta de referência;

(2) dar apoio a professores e alunos ao criar estruturas conceituais na elaboração de materiais didáticos, aprimorando a comunicação do conhecimento de um dado domínio e, assim, o seu aprendizado;

(3) apoiar a implantação de projetos de pesquisa ou de atividades profissionais ao criar uma base de conhecimento de auxílio à criação de um contexto conceitual de estudo;

(4) proporcionar classificações para diferentes finalidades, tais como classificação de doenças e de competências para atribuição de tarefas;

(5) oferecer uma base de conhecimento para a construção de mecanismos de buscas de apoio à recuperação de informação;

(6) auxiliar o desenvolvimento de *software* ao fornecer uma base conceitual para a definição de elementos de dados e de hierarquias de objetos.

Os diversos tipos de SOCs possuem distintos níveis de controle terminológico de um determinado domínio (campo do saber ou assunto, atividade corrente ou tarefa). Eles também são construídos com diferentes abordagens de modelagem, sobretudo no que diz respeito à indicação de relacionamentos entre os conceitos de sua estrutura.

Segundo Hodge (2000), os diversos tipos de SOCs podem ser sistematizados da seguinte maneira: (1) grupo de instrumentos compostos por listas de termos: arquivo de autoridade, glossários, gazetteers, e dicionários; (2) grupo de instrumentos compostos por classificações e categorizações: lista de cabeçalhos de assunto, sistemas de classificação bibliográfica, taxonomias e sistemas de classificação bibliográfica facetados; (3) grupo de instrumentos compostos por listas de termos e relacionamentos: tesauros, redes semânticas e ontologias.

Entre os diferentes tipos de SOCs, os tesauros são linguagens de indexação, construídos a partir de um conjunto de regras pré-estabelecidas e constituídos por descritores preferidos e não-preferidos, que representam conceitos que podem ser combinados no momento de seu uso (pós-coordenação) e usualmente são restritos a uma única especialidade.

Em geral, os tesauros são apresentados na forma alfabética e na forma sistemática, que oferece elementos de significação, permitindo ao usuário a apreensão do conhecimento de um domínio por meio das relações estabelecidas entre conceitos. Tradicionalmente, eles têm dois planos de trabalho: o plano das ideias e o plano verbal (TRISTÃO; FACHIN; ALARCON, 2004). Alguns tesauros também podem oferecer o plano notacional (tal como o existente nos sistemas de classificações), o que possibilita a localização de recursos informacionais.

Os projetos de construção de tesauros possuem três etapas básicas: (1) inicial: composição de uma equipe de trabalho; planejamento; resolução sobre os objetivos; definição do público-alvo; levantamento de terminologia; (2) desenvolvimento: concepção da estrutura conceitual; compilação e seleção do conjunto de termos; definições dos conceitos; determinação dos descritores (preferidos e não-preferidos); agrupamentos em classes básicas e facetas; elaboração dos de mapa conceitual; atribuição de relações entre conceitos; (3) edição: construção da estrutura conceitual; seleção de *software* para edição do tesauro; elaboração de notas de escopo; determinação da forma de apresentação.

Assim, ao construir um tesauro, cria-se um sistema de conceitos que é composto por um conjunto de conceitos relacionados semântica e genericamente entre si, permitindo diferentes tipos de organização (por exemplo: alfabética, relacional, estruturada por campos semânticos, entre outros). Nesse sentido, Currás (1995) afirma que em ambiente organizacional os tesauros desempenham as funções de representação de assuntos e de apoio às consultas de busca dos usuários, auxiliando o processo de recuperação da informação.

Tradicionalmente, os tesauros possuem uma estrutura semântica constituída por uma rede de três distintos tipos de relações: (1) de equivalência, com controle de termos em sinonímia e controle de variações linguísticas; (2) hierárquicas, com agrupamentos constituídos por conceitos ordenados em níveis diferentes de generalidade e de especificidade; e (3) associativas, com a atribuição de ligações não-hierárquicas entre conceitos.

Essa estrutura conceitual dos tesauros vem evoluindo (MOTTA, 1987; CAMPOS, 1995; CAMPOS; GOMES, 2003) e já há evidências de diferentes desdobramentos para cada um desses tipos de relações. Como, por exemplo, no caso da relação de equivalência, ela pode variar desde uma equivalência ortográfica, total ou parcial, até ao uso de abreviaturas, nome fantasia ou equivalência em outro idioma. Essa situação se repete, também, para os relacionamentos hierárquicos e associativos.

Estruturalmente, um tesauro é também composto por: (1) uma terminologia (descritores preferidos e não-preferidos); (2) uma estrutura gramatical (forma de apresentação e de composição dos descritores); (3) uma rede paradigmática (*a priori*); (4) uma rede sintagmática (*a posteriori*). O conjunto desses elementos tem como consequência a não existência de qualquer descritor em um tesauro sem que esteja diretamente relacionado à significação de outro descritor na sua estrutura (Svenonius, 2000). Entretanto, não se pode deixar de considerar que a construção de tesauros deve ser sempre dependente do propósito de sua elaboração, de seu uso e do domínio modelado.

A estrutura conceitual do tesauro é modelada já visando a minimizar a ambiguidade (imprecisão do significado) e a polissemia (pluralidade de significados) da linguagem natural. Apesar de a estrutura conceitual do tesauro possuir essa semântica bastante rica, a falta de especificação dos distintos tipos de relacionamentos existentes entre os conceitos e termos ainda pode ser considerada uma limitação para o uso do tesauro em ambiente digital e para que ele seja classificado como um tipo de SOC.

Nesse contexto, o objetivo deste estudo foi aplicar um modelo de reengenharia de tesauro tradicional para torná-lo um instrumento mais formalizado, de tal forma que os relacionamentos semânticos entre os conceitos do sistema estejam identificados e explicitados para o usuário.

2. Fundamentos teóricos

As bases teóricas utilizadas neste estudo deram subsídios para o seu desenvolvimento e para a aplicação do modelo de reengenharia escolhido. Foram feitas descrições e reflexões sobre os fundamentos empregados na construção de tesauros, a partir da literatura da BCI, Terminologia, Pragmática e Semântica.

Da BCI foram utilizados os princípios teóricos para a organização e sistematização dos conceitos de um domínio, sobretudo com as bases da Teoria da Classificação Facetada, de Ranganathan (1967) e a Teoria do Conceito, de Dahlberg (1978). Na primeira teoria, Ranganathan apresenta o conceito como uma unidade do pensamento,

> [...] um corpo de ideias organizado ou sistematizado, cujas extensão e intensão devem ser coerentes com o domínio de interesse e ajustadas à competência intelectual e ao campo especializado de qualquer indivíduo (RANGANATHAN, 1967, p. 82).

Para Ranganathan, o conceito é concebido no plano das ideias, por meio de distintos procedimentos, a saber:

> (1) processo de definição do assunto; (2) seleção das características que constituem o assunto; (3) seleção de um modelo para o mapeamento da informação sobre os conceitos; (4) agrupamento e divisão destes conceitos conforme suas características comuns e diferentes; (5) organização e o arranjo de grupos e subgrupos (LIMA, 2007, p. 32).

O conjunto de procedimentos se refere ao processo de análise conceitual e, ao final, há a reprodução mental do objeto e a sua representação na forma de um termo (RANGANATHAN, 1967). Kobashi e Francelin (2011) alegam que a

> [...] lógica subjacente à sua teoria indica que o conceito se estabelece em uma sequência de etapas, recortes, enfim, 'fatias' que determinam o movimento contínuo e infinito de sua Espiral do Universo do Conhecimento (KOBASHI; FRANCELIN, 2011, p. 10).

Para as autoras, as relações entre termos e conceitos são estabelecidas nessa espiral, cuja estrutura conceitual criada representa percepções individuais sobre o mundo real. Sobre Ranganathan, Campos e Gomes (2003) afirmam que

> Ranganathan elabora uma série de princípios que visam a permitir que os conceitos de um domínio de saber possam ser estruturados de forma sistêmica, isto é, os conceitos se organizam em renques e cadeias, essas estruturadas em classes abrangentes, que são as facetas, e estas últimas dentro de uma dada categoria fundamental. A reunião de todas as categorias forma um sistema de

> conceitos de uma dada área de assunto e cada conceito no interior da categoria é também a manifestação dessa categoria (CAMPOS; GOMES, 2003, p. 158).

Com esses princípios, Ranganathan (1957) aponta que na espiral há diversos tipos de relacionamentos entre conceitos e termos, o que torna possível também representar realidades complexas (sentidos multidirecional e multidimensional). Com isso, o conhecimento de um domínio é representado como um organismo vivo (metáfora da árvore Baniana), que é dinâmico, pois está em constante desenvolvimento e suas unidades de conhecimento se inter-relacionam entre si.

Sobre a Teoria do Conceito, Dahlberg (1992) afirma que ela compõe o campo da Organização do Conhecimento, juntamente com os campos filosóficos da Lógica, Teoria da Ciência, Epistemologia, Ontologia, Fenomenologia, Aletiologia e Metafísica.

Dahlberg (1978a) desenvolveu a sua teoria tendo por base os princípios analíticos de Aristóteles e os princípios analítico-sintéticos da teoria de Ranganathan. Os princípios analíticos auxiliam a fatoração do objeto representado em suas partes constituintes (elementos individuais), determinando uma hierarquia de proposições verdadeiras sobre o objeto, dando origem a uma classe mais genérica. Já os princípios analítico-sintéticos permitem a integração desses elementos sistêmicos (proposições) que culminam na identificação do termo que representa o conceito. Assim, na Teoria do Conceito, o

> conceito é uma unidade do conhecimento, compreendendo afirmações verdadeiras sobre um dado item de referência, representado numa forma verbal [sendo que:] afirmação verdadeira é a componente de um conceito que expressa um atributo do seu item de referência; item de referência é o componente de um conceito para o qual sua afirmação verdadeira e sua forma verbal estão diretamente relacionadas, sendo assim seu 'referente'; forma verbal (termo/nome) de um conceito é o componente que resume convenientemente ou sintetiza e representa um conceito com o propósito de designar um conceito em comunicação (DAHLBERG, 1978b, p. 147, grifos da autora citada).

Dessa forma, nota-se que para Dahlberg (1978) o conceito é formado pela tríade: (1) referente (objeto a ser conceitualizado); (2) características (todos os enunciados verdadeiros a respeito do referente); (3) forma verbal (termo), conforme Figura 1.

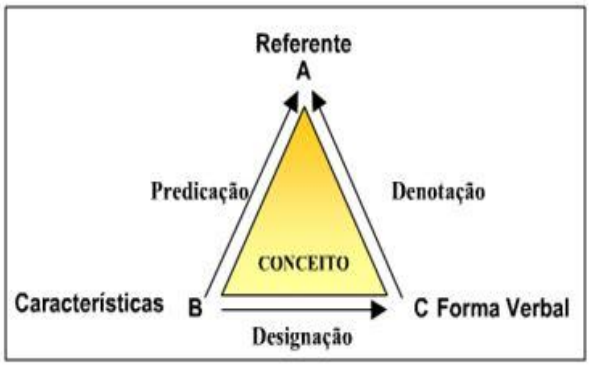

Figura 1. *Triângulo do Conceito (Dahlberg, 1978b, p. 149).*

Assim, o objeto (referente) é fatorado por predicações (características) que são as afirmações verdadeiras sobre esse objeto, cujo conjunto será designado por uma forma verbal (ou outro símbolo). A teoria desenvolvida por Dahlberg possui forte natureza analítica e lógico-positivista, pois utiliza uma abordagem na qual se determina a univocidade do significado de um termo.

Essa natureza lógico-positivista da Teoria do Conceito tem sua origem no campo da Terminologia, advinda dos princípios normativos desenvolvidos por Wüster na Teoria Geral da Terminologia (TGT). Em seus estudos, Eugen Wüster, fundador da Escola Terminológica de Viena, criou métodos para a compilação, padronização e organização da terminologia da área da eletrotécnica, que tinha como objetivo permitir a comunicação entre os profissionais dessa área (KRIEGER; FINATTO, 2004). A TGT pode ser sintetizada pelas características:

> a) a prioridade do conceito em detrimento do termo; b) a precisão do conceito, o que retoma, de certo modo, a eliminação da ambiguidade e a busca da univocidade; c) a consequente abordagem onomasiológica, já que toda a atividade terminológica parte do conceito; d) a proeminência do nível lexical em detrimento dos demais níveis de descrição linguística (morfológico, sintático, textual, discursivo); e finalmente; e) a prescrição (ALMEIDA, 2006, p. 86).

Nessa perspectiva, esse conjunto de características torna o produto terminológico bastante rígido, o que pode ser elucidado porque a TGT

> [...] tem como foco principal o componente conceitual, em detrimento do significado [...] os conceitos, nesta perspectiva, são estáveis, paradigmáticos e universais, como idealiza o lógico-positivismo (KAMIKAWACHI, 2010, p. 21).

Cabré (1999) aponta que esse caráter prescritivo e normativo pode ser válido em alguns contextos de controle terminológico. Porém, a partir

de 1990 houve a expansão das reflexões sobre a área da Terminologia e críticas às bases teóricas da TGT. Com isso, novas abordagens surgiram e, entre elas, a Teoria Comunicativa da Terminologia (TCT), desenvolvida por Cabré.

As bases da TCT estão fundamentadas em três teorias: (1) do conhecimento, de natureza cognitiva, para o entendimento sobre os conceitos, suas designações e as inter-relações estabelecidas entre eles; (2) da comunicação, que estabelecem situações comunicativas, assim como as suas características, perspectiva, propósito e limitações; e (3) da linguagem, que permite compreender as unidades léxicas, na linguagem comum e de especialidade, e os seus contextos de uso (Cabré, 1999).

Com a combinação dos princípios dessas três teorias é possível perceber as unidades terminológicas a partir de seus aspectos linguísticos e comunicativos, que têm comportamento semelhante às palavras do léxico de uma língua geral. Para Cabré (1999), a linguagem de especialidade é regida pelas mesmas regras e é caracterizada pelos mesmos fenômenos de sinonímia e variação linguística, presentes na linguagem geral.

Portanto, o que difere um termo de uma palavra é, principalmente, porque os termos que compõem uma linguagem especializada possuem características específicas que os tornam termos advindos de uma temática única, utilizados por um grupo específico de interlocutores e em um contexto de comunicação. Assim, "um termo é uma unidade linguística que tem uma função comunicativa e pragmática" (Maculan, 2015, p. 74).

As teorias da Pragmática proporcionaram as bases para a exploração de relações mais contextualizadas ao ambiente de uso. Segundo os princípios da Pragmática, não existe uma verdade absoluta, que possa ser generalizada a toda situação de uso. As verdades são um construto social de sentido e estão sujeitas às mudanças no tempo e no espaço. Assim, a verdade é estabelecida para

> [...] dar conta especificamente da consideração da linguagem como ação, como produzindo efeitos e consequências em contextos determinados (MARCONDES, 2000, p. 41).

Nesse sentido, cada verdade é validada a partir dos efeitos práticos que é capaz de produzir para atender a uma comunidade de usuários ou domínio. Essa verdade é representada por meio de uma linguagem que, segundo Wittgenstein, é compartilhada por uma comunidade discursiva, mas não própria de um único indivíduo. Dessa forma,

> [...] a importância do uso ganha uma dimensão mais complexa em Wittgenstein porque não se refere apenas à inserção de palavras em frases, mas a uma situação de ação com finalidade prática, como um exercício de influência de uns sobre os outros em um ambiente complexo. A esse ambiente, o autor denominou "jogo de linguagem" ou Sprachspiel, uma atividade regulada e partilhada (SOUZA; HINTZE, 2010, p. 115).

Nota-se, assim, que o "jogo" é composto por regras que devem ser seguidas entre os sujeitos, a partir de um consenso no uso dessa linguagem, visando à comunicação.

Da Semântica foram aplicados os princípios dos campos semânticos, uma vez que os conceitos, termos e relações têm suas significações dependentes do valor que lhes são impressos.

Para criar a taxonomia dos relacionamentos foi preciso entender o valor semântico que os verbos abarcam no português brasileiro. Os verbos carregam um valor semântico, e compreender esse valor é importante, conforme é salientado por Soergel *et al.* (2004). Para os autores, os relacionamentos em um tesauro devem ser representados por expressões verbais. Para isso, buscamos respaldo em duas teorias: na Teoria da Valência (TV), criada por Francisco S. Borba (1996), e em parte da Teoria do Léxico Gerativo, desenvolvida por Pustejovsky (1995), especificamente do elemento que esclarece sobre a definição dos papéis *Qualia* para os verbos.

Borba (1996, p. xxi) conceitualiza a TV como um "conjunto de relações estabelecidas entre o verbo e seus argumentos ou constituintes indispensáveis". Essa teoria foi desenvolvida a partir da combinação de duas outras teorias: a Gramática de Valências (Tesnière, 1966; Chafe, 1970; Vilela, 1992) e a Gramática de Casos (Fillmore, 1968, 1969, 1977; Anderson, 1971; Jackendoff, 1972; Cook, 1979, 1989).

Para Borba, a valência de um verbo representa o número de argumentos que ele necessita preencher para completar seu sentido e essa valência pode ser classificada em três níveis:

(1) Valência quantitativa (lógica ou lógico-semântica): número de argumentos que um predicado pode ter: avalente (zero), monovalente (um), divalente (dois), trivalente (três) e tetravalente (quatro);

(2) Valência qualitativa (sintática ou morfossintática): características dos actantes (relações gramaticais e/ou funcionais), das propriedades morfológicas, das funções sintáticas, das pro-

priedades sintáticas e das classes que preenchem os argumentos;

(3) Valência semântica: traços semânticos das categorias (+humano, +animado, +contável), das funções ou dos papéis temáticos (agente, causativo, beneficiário) e das restrições relacionais de coocorrência ou exclusão.

Para cada situação específica, em um dado domínio, é possível determinar qual a classe dos verbos (ação, processo, atividade e estado). Assim, a ausência de um agente permite que se distinga um verbo de processo de um verbo de ação, ou um verbo de atividade de um verbo de estado (Schwarze, 2001, p. 97).

Os verbos podem exprimir diversas operações, que se referem às funções ou propósitos de objetos reais, imaginários ou abstratos. Dada essa multiplicidade de dimensões de significados dos verbos, a estrutura dos papéis *Qualia*, que é parte da Teoria do Léxico Gerativo, pode ser utilizada como passo inicial na identificação do significado semântico de um verbo. A determinação dos papéis *Qualia* para as expressões verbais, que representam as relações semânticas na estrutura de um tesauro, permite identificar os fenômenos de hiperonímia/hiponímia (relação de gênero-espécie) e de holonímia/meronímia (relação todo-parte).

A estrutura *Qualia* é composta pelos papéis Formal, Constitutivo, Télico e Agentivo, conforme descrito a seguir:

(a) Formal: generalização de uma operação descrita através de outra operação representada; distingue um objeto em um domínio mais amplo ou geral;

(b) Constitutivo: constituição de uma operação descrita, expondo-a por meio de outras operações que são necessárias para efetivá-la; indica uma relação entre um objeto e suas partes constituintes;

(c) Agentivo: especificação da entrada de uma operação, na forma de argumentos, representados por objetos reais, imaginários ou abstratos; indica elementos ou fatores que estão envolvidos na origem do objeto ou as causas para o objeto acontecer, existir ou ocorrer;

(d) Télico: especificação da saída de uma operação, na forma de objetos reais, imaginários ou abstratos; expressa o propósito e a função do objeto.

Tanto o sentido verbal (entre o verbo e o sujeito) quanto o sentido nominal (entre o substantivo e o adjetivo) dos papéis *Qualia* estabelecem as relações existentes entre os termos em um determinado contexto comunicativo. Assim, os papéis *Qualia* que interessam ser representados no refinamento das relações semânticas em tesauros são os sentidos verbais das relações que expressam a ligação que existe entre dois conceitos em um dado domínio.

3. Metodologia

O modelo de reengenharia de tesauros tradicionais estudado foi desenvolvido por Soergel *et al.* (2004) e, depois, detalhado por Lauser *et al.*(2006), cuja estrutura conceitual pode ser observada na Figura 2.

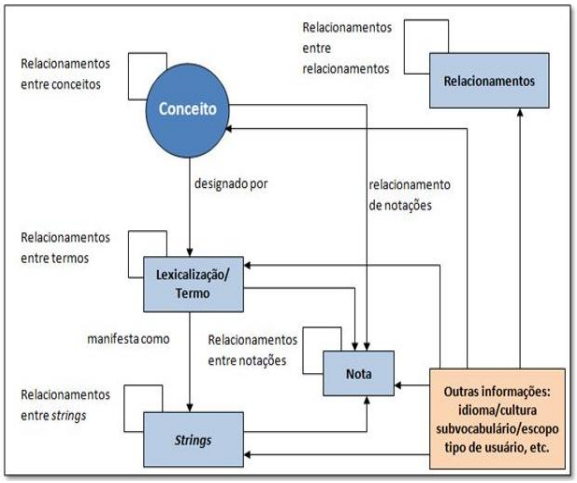

Figura 2. *Modelo para reengenharia de tesauro*

Observando a Figura 2 é possível perceber que os autores propuseram um modelo de reengenharia de tesauro cuja principal característica é a individualização da modelagem em cinco distintos níveis de entidades: conceito, termo ou lexicalização, *string* ou variações linguísticas, notas de escopo e relacionamentos. Com isso, os relacionamentos podem ocorrer entre entidades de mesmo tipo e entre entidades de tipos diferentes. Ademais, uma importante particularidade do modelo é a distinção entre os termos que designam os conceitos e as manifestações desses termos, representadas pelos strings, que se referem às variações linguísticas (singular/plural; variação regional).

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)

Figura 3. *Estrutura relacional do Thesagro*

Dessa maneira, os relacionamentos são a espinha dorsal do tesauro.

Antes da aplicação do modelo de reengenharia de tesauro, houve um planejamento inicial no qual foi selecionada uma equipe de trabalho, o domínio modelado, da Agropecuária, e selecionou o tesauro *Thesaurus* Agrícola Nacional (Thesagro), único vocabulário controlado em português brasileiro, para a aplicação do modelo de reengenharia de tesauro.

No percurso metodológico foram utilizadas duas ferramentas: o sistema Termos Eletrônicos (e-Termos) e o Extrator de Termos e Estruturas Conceituais Agrícolas Multilíngue (ETECAM).

O sistema e-Termos é um ambiente computacional colaborativo *web*, de utilização gratuita e restrita aos usuários cadastrados. A ferramenta é composta por seis etapas, possuindo um conjunto de procedimentos automatizados e semi-automatizados, que têm como objetivo dar suporte à criação e gestão de produtos terminológicos para distintos fins (ensino, glossários, vocabulários controlados). O sistema e-Termos tem por base os fundamentos teóricos da Teoria Comunicativa da Terminologia (TCT), desenvolvida por Cabré (1999), que foi adotada no desenvolvimento deste estudo.

A ferramenta ETECAM foi criada para uso privativo da Embrapa, sendo utilizada para executar uma comparação automática sobre a existência de termos coincidentes entre terminologias de distintos tesauros. Essa ferramenta realiza as tarefas: (a) entrada de uma lista de termos, (b) a ferramenta verifica a existência de termos coincidentes à lista, (c) extrai os termos e seu *cluster* semântico e (d) os equivalentes em outros idiomas.

4. Aplicação das etapas e resultados

O modelo escolhido para a reengenharia é composto por três etapas básicas: (1) definição da estrutura do tesauro, (2) coleta de terminologia e (3) edição do tesauro.

4.1. Primeira etapa

A definição da estrutura do tesauro envolveu o mapeamento das características estruturais do tesauro Thesagro e a determinação como recorte temático a subárea da Intensificação Agropecuária.

O conceito de Intensificação Agropecuária adotado neste estudo segue a teoria de Boserup (1965), ou seja, é a relação entre o crescimento ou não da população de uma região e a determinação do "aumento da produção total agrícola em uma mesma unidade de área ou, ainda, a manutenção de certa produção com uma menor quantidade de insumos" (OLIVEIRA, 2011, p. 4).

O Thesagro possui cerca de 9.400 descritores, todos identificados por um ID (identificador) numérico.

A sua estrutura conceitual é composta pelos três relacionamentos básicos de qualquer tesauro tradicional, ou seja, relações de equivalência, hierárquicas e associativas.

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

Figura 4. *Estrutura relacional do Thesagro*

Nota-se que os símbolos dos diferentes tipos de relacionamentos são representados na língua inglesa.

Segundo Maculan (2015):

> Foram identificadas cerca de 12.000 representações de relações associativas (RT). A versão impressa do THESAGRO apresenta uma nota explicativa que esclarece e justifica o uso de um alto número de ligações associativas (RT): para os responsáveis pela manutenção do THESAGRO, essa foi uma decisão intencional (Maculan, 2015, p. 200).

O objetivo desse tipo de abordagem é manter associações entre conceitos que não são intuitivos de serem percebidos, principalmente pelos usuário não especialista.

A modelagem da subárea da Intensificação Agropecuária teve como ponto de partida uma taxonomia já existente, elaborada pelos especialistas da Embrapa, que foi composta por 639 conceitos da temática.

Essa taxonomia foi estruturada em nove classes básicas: (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.

Uma vez que a taxonomia inicial era muito extensa, houve a necessidade de aplicar a modelagem de reengenharia do Thesagro usando uma amostra, conforme Figura 5.

1. INTENSIFICAÇÃO AGROPECUÁRIA 2. AGRICULTURA EXTENSIVA 3. AGRICULTURA INTENSIVA	MATERIAL E MÉTODOS 4. cultura 5. sensoriamento remoto 6. sistema de informação geográfica
AMBIENTE 7. meio ambiente 8. ambiente físico 9. solo	AGRONOMIA 10. manejo da cultura 11. manejo do solo 12. pousio 13. período de pousio 14. pesticida 15. ciclo da cultura 16. cultura anual 17. pecuária 18. adubo verde 19. produto agropecuário 20. biomassa
TERRITÓRIO E PAISAGEM 21. posse da terra 22. escassez de terra 23. cobertura da terra 24. mudança de cobertura da terra	
SOCIOECONOMIA 25. densidade demográfica 26. crescimento populacional 27. pressão populacional	ESPAÇO E TEMPO 28. mudança agrícola 29. análise de séries temporais
INSTITUIÇÕES 30. Embrapa	

Figura 5. *Amostra de estudo e análise*

Assim, a validação do modelo foi realizada pela modelagem e análise de uma amostragem intencional, composta por 30 conceitos representativos do conjunto de classes básicas e da área temática escolhida.

4.2. Segunda etapa

A coleta de terminologia teve como insumos terminológicos a taxonomia da Intensificação Agropecuária e outros três tesauros: o Thesagro, o Agrovoc e o *National Agricultural Library* (NAL).

Nessa atividade, ocorreu a comparação entre os conceitos da amostra, oriundos da terminologia da taxonomia, e a terminologia existente em cada um dos três tesauros selecionados. Para essa comparação, a listagem original dos conceitos da amostra foi subdividida em duas listas:

Lista Um: composta pelos 30 conceitos da amostra, em português brasileiro, adicionando-se as expressões desses termos no singular e plural;

Lista Dois: composta pelos 30 conceitos da amostra, traduzidos para o inglês, adicionando-se as expressões desses termos no singular e plural, assim como na sua forma inversa (adjetivo + substantivo), por essa inversão ser comum na língua inglesa.

O procedimento da comparação terminológica foi realizado com a ferramenta ETECAM, que permitiu recuperar os termos coincidentes com as duas listas, comparando os termos de cada uma delas com a terminologia existente nos três tesauros, separadamente. Depois dessa comparação, foi realizada também uma comparação intelectual, possibilitando adicionar os conceitos e seus *clusters*, que não haviam sido resgatados com a comparação automática.

4.3. Terceira etapa

A edição do tesauro Thesagro incluiu a atividade intelectual da modelagem do recorte temático da Intensificação Agropecuária e a inserção da estrutura conceitual no sistema e-Termos, para a gestão da terminologia. Assim, essa etapa foi constituída pelos seguintes procedimentos: compilação da base definicional, elaboração de glossário, confecção das fichas terminológicas, determinação de notas de escopo e construção do sistema de conceitos.

A compilação da base definicional foi realizada de forma constante e dinâmica, e recolheu e armazenou, no sistema e-Termos, contextos explicativos e/ou definitórios sobre o domínio modelado. Esse recurso informacional auxiliou a elaboração das definições terminológicas dos conceitos da amostra, dando origem ao glossário.

Para cada um dos conceitos da amostra foi confeccionada uma ficha terminológica, composta por 38 campos semânticos, para preenchimento. Dentre esses campos se destacam: definições dos conceitos (do especialista, modelador e final), informações enciclopédicas e de glosa, notas de escopo, termos em relação de equivalência, de variação linguística, assim como os conceitos em relação hierárquica e associativa. Destaca-se que foram elaboradas notas de escopo para alguns dos conceitos da amostra, a partir da avaliação feita pela equipe de trabalho quanto à necessidade dessa nota explicativa.

A construção do sistema de conceitos foi realizada a partir dos conteúdos das definições e dos registros das fichas terminológicas. Essa atividade envolveu a aplicação de 44 diferentes relações, que criou uma rede semântica para os conceitos da amostra e seus *clusters* semânticos, gerando um desdobramento que totalizou cerca de 600 relacionamentos.

Foram representados os relacionamentos: de gênero e suas espécies, do todo e suas partes, de equivalências, de *strings* (variações) e associativas.

Para exemplificar e comparar a atual estrutura do Thesagro e a modelagem realizada utilizan-

do o modelo de reengenharia de tesauro, a seguir apresenta-se a atual modelagem do descritor GATO no Thesagro e, depois a estrutura refinada:

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

Figura 6. *Estrutura atual do Thesagro*

Na estrutura atual do Thesagro, percebe-se que os relacionamentos são ainda bastante genéricos. Além disso, os descritores FELIS CATUS DOMESTICUS e FELIS DOMESTICA são nomes científicos para GATO, mas essa informação fica perdida na estrutura atual do Thesagro, pois os descritores estão ligados por uma relação associativa e não por uma relação de equivalência.

Aplicando o modelo de reengenharia 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

Figura 7. *Estrutura reformulada do Thesagro*

Nota-se que houve o refinamento das relações estabelecidas entre os conceitos e termos, com a explicitação do tipo de ligação há entre eles, facilitando o entendimento da estrutura semântica. Com a reformulação da estrutura do Thesagro ficou claro ao usuário quais são os nomes científicos utilizados para representar o conceito GATO, que é o nome popular desse animal.

Outro exemplo é a estrutura original do Thesagro para o descritor PESTICIDA e a estrutura remodelada, utilizando o modelo de reengenharia de tesauros:

(1) ESTRUTURA ATUAL	(2) REENGENHARIA DA ESTRUTURA
PESTICIDA *NT* ACARICIDA *NT* CARRAPATICIDA *NT* FUNGICIDA *NT* GERMICIDA *NT* INSETICIDA *NT* MOLUSCICIDA *NT* NEMATICIDA *NT* PERSISTÊNCIA DE PESTICIDA *NT* RATICIDA *NT* REPELENTE	PESTICIDA *termoGenéricoGênero (TGG)* SUBSTÂNCIA QUÍMICA *termoEspecíficoGênero (TEG)* ACARICIDA *termoEspecíficoGênero (TEG)* CARRAPATICIDA *termoEspecíficoGênero (TEG)* FUNGICIDA *termoEspecíficoGênero (TEG)* GERMICIDA *termoEspecíficoGênero (TEG)* INSETICIDA *termoEspecíficoGênero (TEG)* MOLUSCICIDA *termoEspecíficoGênero (TEG)* NEMATICIDA *termoEspecíficoGênero (TEG)* RATICIDA *termoEspecíficoGênero (TEG)* REPELENTE *temPropriedade (TR)* PERSISTÊNCIA DE PESTICIDA

Figura 8. *Estrutura conceitual: descritor PESTICIDA*

É possível notar que na estrutura original (1) do THESAGRO o descritor PESTICIDA não está ligado a um conceito superordenado (hiperônimo), que indique o seu pertencimento a uma classe mais geral. Também observamos as relações hierárquicas entre PESTICIDA e os outros 10 descritores, mas são ligações genéricas que não identificam se são relações do tipo gênero-espécie, todo-parte ou de instância.

Ao aplicar a reengenharia (2) nessa estrutura do Thesagro, atribuímos uma relação hierárquica, do tipo gênero-espécie, entre PESTICIDA (TEG) e SUBSTÂNCIA QUÍMICA (TGG), que indica a classe mais geral de PESTICIDA. Com exceção do descritor PERSISTÊNCIA DE PESTICIDA, foram refinadas as relações hierárquicas entre PESTICIDA e os outros descritores, identificadas como relações do tipo gênero-espécie (os conceitos subordinados carregam as mesmas características do conceito PESTICIDA).

Com o descritor PERSISTÊNCIA DE PESTICIDA foi percebido um problema de abstração conceitual na estrutura original do Thesagro, pois ele não é um tipo de PESTICIDA, mas uma propriedade que indica o período durante o qual a toxicidade do pesticida permanece inalterada (longevidade do produto), afetando o ambiente no qual foi aplicado. Assim, criamos uma relação associativa, do tipo <tem_propriedade>, entre PESTICIDA e PERSISTÊNCIA DE PESTICIDA, facilitando a compreensão do domínio.

Na representação e refinamento dos relacionamentos na estrutura conceitual dos conceitos da amostra, houve predominância dos relacionamentos hierárquicos, totalizando 286 ocorrências, 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). Esse resultado demonstrou, na nova estrutura semântica construída para o Thesagro, a manutenção da sua natureza específica de origem.

Quanto às relações associativas, foram representados 232 relacionamentos, evidenciando a complexidade da subárea da Intensificação Agropecuária.

Os resultados demonstraram que a expressão explícita das relações entre pares de entidades (conceitos, termos, strings e notas de escopo) refinou a semântica da estrutura do tesauro, dando subsídios para facilitar a interoperabilidade entre diferentes tesauros ou sistemas.

5. Considerações finais

Este estudo teve como ambientação a Embrapa Informática Agropecuária (Embrapa), Unidade Campinas/SP, sendo o primeiro resultado do convênio firmado entre essa instituição, a Universidade Federal de Minas Gerais (UFMG) e o Grupo de Pesquisa Protótipo Mapa Hipertextual (MHTX).

O desenvolvimento deste estudo partiu do pressuposto de que os tesauros tradicionais já possuem uma representação semântica bastante consistente, mas ainda insuficiente para classificá-lo como um SOC, que abarca instrumentos que podem ser entendidos pela máquina.

Dessa maneira, neste estudo foi priorizada a explicitação dos relacionamentos entre conceitos e termos, que torna a estrutura conceitual do tesauro semanticamente mais rica.

Ao se desenvolver uma nova estrutura semântica para o Thesagro, ficou evidente a necessidade da adoção de normas e padrões internacionais na criação de tesauros, que oferece parâmetros que permitem maior formalidade nas representações dos relacionamentos, auxiliando a interoperabilidade com outros vocabulários.

Nesse sentido, o uso do modelo de dados *Simple Knowledge Organization Systems* (SKOS) possibilita a explicitação de relações que podem ser interpretadas (legíveis) por máquinas, sendo um elemento importante que pode auxiliar na interoperabilidade entre diferentes vocabulários e sistemas.

Por fim, a aplicação do modelo de reengenharia selecionado auxiliou a organização e a ampliação da visão acerca do conhecimento da área modelada, o que poderá facilitar a sua compreensão e exploração pelo usuário, uma vez que tal refinamento os tornou mais claros e específicos.

Referências

Anderson, J. M. (1971). *The grammar of case:* towards a localist theory. London: CUP.

Borba, F. S. (1996). *Uma gramática de valências para o português*. São Paulo: Ática.

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

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

Campos, M. L. A. (1995). Linguagens documentárias: núcleo básico de conhecimento para seu estudo. *R. Esc. Biblioteconomia UFMG*, Belo Horizonte, 24:1, (jan./jun.) 52-62.

Campos, M. L. A.; Gomes, H. E. (2003). Organização de domínios de conhecimento e os princípios ranganathianos. *Perspectivas em Ciência da Informação*, Belo Horizonte, 8:2 (jul./dez.).

Chafe, W. L.(1970). *Meaning and the structure of language*. Chicago: University of Chicago.

Cook, W. A. S. J. (1979). *Case grammar:* development of the matrix model (1970-1978). Washington, D.C.: Georgetown University.

Cook, W. A. S. J.(1989). *Case grammar theory model*. Washington, D.C.: Georgetown University.

Dahlberg, I. (1978a). Teoria do conceito. Tradução Astério Tavares Campos. *Ciência da Informação*, Rio de Janeiro, 7:2, 101-107.

Dahlberg, I. (1978b). A referent-oriented, analytical concept theory of Interconcept. *International Classification*, 5:3, 122-151.

Dahlberg, I. (1992). Knowledge organization and terminology; philosophical and linguistic bases. *International Classification*, 19:2, 65-71.

Fillmore, C. J. (1968). The case for case. In: BACH, E.; HARMS, R.T. (Org.) *Universals in linguistic theory.* New York: Rinehard and Winston. p. 1-88.

Fillmore, C. J. (1969). Types of lexical information. In: KIEFER, F. (Ed.). *Studies in syntax and semantics*. Dortrecht-Holland: D. Reidel.

Fillmore, C. J. (1977). *A semântica na linguística moderna:* o léxico. Tradução de Lúcia M. Lobato. Rio de Janeiro: Francisco Alves.

Hodge, G. (2000). Systems of knowledge organization for digital libraries: beyond traditional authorities files. Washington, DC: Council on Library and Information Resources. Recuperado 24-06-2013, de http://www.clir.org/pubs/re-ports/pub91/contents.htm.

Jackendoff, R. (1972). *Semantic interpretation in generative grammar*. Massachusetts: MIT Press, Cambridge.

Kobashi, N. Y.; Francelin, M. M. (2011). Conceitos, categorias e organização do conhecimento. *Informação e Informação,* Londrina, 16:3, (jan./jun.), 1-24.

Lauser, B. *et al*. (2006). *From Agrovoc to the Agricultural Ontology Service: Concept Server an OWL model for creating ontologies in the agricultural domain*. En International Conference on Dublin Core and Metadata Applications, 2006, Colima, Mexico. México: DCMI.

Lima, G. A. B. O. (2007). *A análise facetada na modelagem conceitual para organização hipertextual de documentos acadêmicos*: sua aplicação no protótipo MHTX (mapa hipertextual). *Informação e Sociedade*: Estudos, João Pessoa, 17:1, (jan./abr.), 31-41.

Maculan, B. C. M. S. (2015). *Estudo e aplicação de metodologia para reengenharia de tesauro: remodelagem do THESAGRO*. 345f. Tese (Doutorado) – Universidade Federal de Minas Gerais, Escola de Ciência da Informação, Belo Horizonte, Brasil.

Marcondes, D. (2000). *Filosofia, linguagem e comunicação*. São Paulo: Cortez.

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

Oliveira, J. R. (2011). Sustentabilidade e intensificação produtiva da agricultura familiar: um estudo comparativo entre duas comunidades em Itapejara D'Oeste, Sudoeste do Paraná. *Synergismus Scyentifica*, Universidade Tecnológica Federal do Paraná, Pato Branco, 6:1.

Pustejovsky, J. (1995). *The generative lexicon*. Cambridge, MA: MIT Press.

Ranganathan, S. R. (1967). *Prolegomena to library classification*. Bombay: Asia Publishing House.

Soergel, D. (1999). The rise of ontologies or the reinvention of classification. *Journal of the American Society of Informationl Science*, 50:12, 1119-1120.

Soergel, D. *et al*. (2004). Reengineering thesauri for new applications: the AGROVOC example. *Journal of Digital Information*, 4:4.

Vickery, B. C. (2007). *A note on knowledge organization*. Site Lifeboat for Knowledge Organization. Recuperado 11-06-2013, de http://www.iva.dk/bh/lifeboat_ko/ concepts /Vickery-_a_note_ on_knowledge_organisation.htm.

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

Schwarze, C.(2001). La sémantique do verbe. In: _____. *Introduction à la sémantique lexicale*. Tübingen: Nar. p. 89-113.

Svenonius, E. (2000). *The intellectual foundations of information organization*. Cambridge: The MIT Press.

Tesnière, L. (1966). *Élements de syntaxe structurale*. 2. ed. Paris: Klincksieck.

Tristão, A. M. D.; Fachin, G. R. B.; Alarcon, O. E. (2004). Sistema de classificação facetada e tesauros: instrumentos para organização do conhecimento. *Ciência da Informação*, Brasília, 33:2, (ago), 161-171.

Vilela, M.(1992). *Gramática de valências:* teoria e aplicação. Coimbra: Almedina.