DOI: 10.5433/1981-8920.2017v22n3p377

SEMIAUTOMATIZAÇÃO DE RELAÇÕES EM TESAUROS: UMA PROPOSTA PARA REFINAMENTO DE RELACIONAMENTOS SEMÂNTICOS A PARTIR DO TESAURO AGROVOC

SEMI-AUTOMATIZACIÓN DE RELACIONES EN TESAUROS: UNA PROPUESTA PARA EL REFINAMIENTO DE RELACIONES SEMÁNTICAS A PARTIR DEL TESAURO AGROVOC

Decio Wey Berti Junior*
Dagobert Soergel **
Gercina Ângela de Lima ***
Benildes Coura Moreira dos Santos Maculan****

RESUMO:
Introdução: Os tesauros são ferramentas que contribuem com a recuperação da informação em serviços como bases de dados digitais e bibliotecas digitais. **Objetivo:** Apresentar uma análise quantitativa do refinamento da estrutura semântica do Tesauro AGROVOC, visando um modelo semiautomatizado para refinamento de relacionamentos semânticos em tesauros. **Metodologia:** Utilizaram-se os dados do Tesauro AGROVOC, refinado e representado em modelo SKOS-XL, agregados à análise qualitativa na classificação de conceitos em tipos de entidade e classificação hierárquica de relacionamentos feita por Soergel. Com esta base foi feita a análise quantitativa dos tipos de relacionamento. **Resultados:** O resultado da análise quantitativa mostra que o refinamento do Tesauro AGROVOC ainda não é completo. A maioria dos relacionamentos *related term* parecem estar refinados, mas os relacionamentos hierárquicos (*broader* / *narrower*) não estão. **Conclusões:** Este estudo demonstra que a análise quantitativa desvenda a estrutura do tesauro para indicar áreas onde é possível implementar melhorias.

*Doutorando em Ciência da Informação pela Universidade Federal de Minas Gerais (UFMG). E-mail: deciowbj@gmail.com
**Doutor em Political Science pela Universität Freiburg, Alemanha.
Professor da University of Buffalo (USA). E-mail: dsoergel@buffalo.edu
***Doutora em Ciência da Informação pela Universidade Federal de Minas Gerais (UFMG). Professora da Escola de Ciência da Informação (UFMG). E-mail: limagercina@gmail.com
*****Doutora em Ciência da Informação pela Universidade Federal de Minas Gerais (UFMG). Professora da Escola de Ciência da Informação (UFMG). E-mail: benildes@gmail.com.

Palavras-chave: Sistema de Organização do Conhecimento. Tesauros. Relações semânticas. AGROVOC.

1 INTRODUÇÃO

A Organização do Conhecimento (OC), enquanto um campo de estudo, segundo Café (2011, p. 25), "fundamenta-se essencialmente em análises de cunho semântico. Relações semânticas são estabelecidas por meio da análise das características ou propriedades dos conceitos, as quais permitem identificar diferenças e semelhanças que evidenciam determinados tipos de relacionamentos". Para Café et al. (2014, p. 204), a OC está "pautada na análise de conceitos, seus significados, relações semânticas e delimitações terminológicas, representando, de forma mais próxima possível, um determinado domínio". Considerar esses elementos é essencial para a evolução da *web*, em direção à *web semântica*, que visa modificar a forma como os conteúdos das páginas são organizados de forma mais significativa, com a implantação de sistemas semânticos, a partir de padrões denominados de *Semantic Web family of standards,* que vêm sendo definidos e mantidos pela *World Wide Web Consortium (*W3C*)*.

Na perspectiva da W3C, neste trabalho considera-se a OC "como o processo de modelagem do conhecimento que visa a construção de representações do conhecimento" (BRÄSCHER; CAFÉ, 2010, p.95) de forma aplicada, por meio de Sistemas de Organização do Conhecimento (SOCs). Dessa forma, para atender aos princípios da W3C, diferentes tipos de SOCs (vocabulários, classificações, taxonomias e tesauro) vêm sendo utilizados na implementação de serviços de informação (banco de dados e de bibliotecas digitais) para garantir a correta interpretação semântica de terminologias. Esses serviços potencializam a interoperabilidade semântica de registros, repositórios, esquemas de metadados, registros de *crosswalk* (emulador) e o mapeamento entre vocabulários e ontologias. Para melhorar a qualidade da recuperação de informações heterogêneas em serviços de informação, homem

e máquina precisam entender (homem) e interpretar (máquina) o significado do que está sendo submetido ao sistema e, assim, saber o que poderá ser recuperado. Para isto é indispensável o uso das interfaces e da interatividade. Sem estes dois fundamentos é impossível haver qualquer tipo de relação homem-máquina dentro da Internet. Dessa forma, o resultado de uma busca deveria conceder, por meio de funcionalidades providas no sistema, com resultados melhores, ou seja, mais relevantes, para o acesso à informação centrado no usuário.

Os padrões de tecnologia para *web* semântica contribuem com representações em linguagem *Simple Knowledge Organization System* (SKOS). A codificação SKOS permite representar os elementos de um tesauro usando o *Resource Description Framework* (RDF), de forma que possam ser lidos por aplicações informatizadas de forma interoperável. Dessa maneira, os relacionamentos são identificados e tipificados, agregando refinamento semântico nos relacionamentos, o que facilita a leitura destes por computadores e, assim, contribui para automatizações.

Partindo do aspecto que se refere ao princípio do refinamento dos relacionamentos semânticos na estrutura de SOCs, em especial dos tesauros, este artigo apresenta os resultados de uma análise quantitativa da estrutura semântica do tesauro AGROVOC que, em sentido mais amplo, espera-se que permita ser generalizado para a determinação de regras de relações (*rules as you go*) em tesauros de forma semiautomática. O restante do artigo está organizado como segue: a seção 2 apresenta o que são os SOCs, em especial, focalizando na estrutura dos tesauros e nos relacionamentos (de equivalência, hierárquicos e associativos) que ocorrem em seu sistema conceitual; a seção 3 descreve o tesauro AGROVOC, enfatizando suas características e o que foi feito no que se refere ao refinamento das relações semânticas em sua estrutura; a seção 4 descreve o modelo de Soergel et al. (2004) denominado de *rules as you go*, que é uma metodologia que permite a manutenção em tesauros com a semiautomatização de relacionamentos em sua estrutura; a seção 5 apresenta a metodologia e os procedimentos, incluindo a sua

aplicação e a análise dos dados obtidos; em seguida, a seção 6 traz as considerações finais sobre o trabalho realizado.

2 SISTEMAS DE ORGANIZAÇÃO DO CONHECIMENTO: TESAUROS

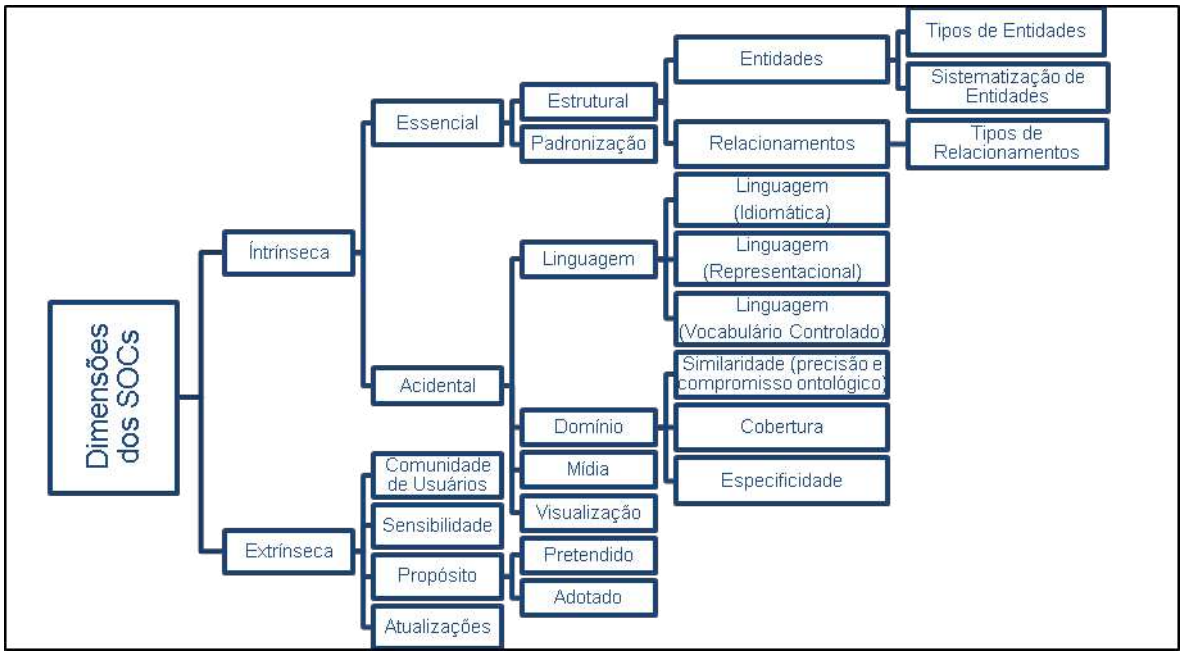
Os sistemas de organização do conhecimento (SOCs) têm por finalidade organizar conteúdos para a recuperação de itens relevantes disponibilizados em bases de dados de uma biblioteca digital. Segundo o autor, tanto as unidades de informação digital quanto as bibliotecas digitais demandam o uso de um ou mais SOCs para prover uma visão geral e a recuperação do conteúdo da coleção. Nesse sentido, os SOCs têm sido utilizados no suporte à Recuperação da Informação (RI) e Aprendizagem de Máquina (AM), e também para auxiliar na atribuição de termos, conceitos e palavras-chave. Eles podem ser utilizados para diferentes finalidades, tais como: (1) prover acesso alternativo a assuntos, (2) adicionar modos de entendimento para recursos de bibliotecas digitais, (3) dar suporte ao acesso multilíngue e (4) fornecer termos para expansão de buscas em texto livre em domínios que são relativamente desconhecidos para o usuário. Eles também podem ser úteis para a categorização automática de documentos, pois pode prover a combinação das técnicas de recuperação da informação e aprendizagem de máquina com um sistema estruturado por especialistas de determinado domínio.

Os SOCs englobam todos os tradicionais instrumentos de organização e gerenciamento do conhecimento e Hodge (2000) oferece uma sistematização que agrupa os SOCs em três grupos de instrumentos: (1) compostos por listas de termos: lista de autoridades, glossários, dicionários e *gazetteers*; (2) compostos por classificações e categorizações: lista de cabeçalhos de assunto, esquemas e classificação bibliográfica, taxonomias e esquemas de categorização bibliográfica facetados; (3) compostos por lista de termos e relacionamentos: tesauros, redes semânticas e ontologias. Nota-se, assim, que os SOCs promovem “níveis diferenciados de estruturação e de controle terminológico” (MACULAN, 2015, p.126), e os diversos tipos têm por base

algumas características tais como estrutura, relacionamento entre termos, função e complexidade.

Sobre esse ponto, Souza, Tudhope e Almeida (2012) acrescentam que um exemplar de qualquer um dos instrumentos pode apresentar diferentes complexidades semânticas, dependendo de sua finalidade e abordagem de criação, o que pode dar origem à sobreposição de características entre os instrumentos. Para os autores, na proposta de Hodge percebe-se um percurso linear de sistematização, sem considerar os distintos níveis de estruturação e, assim, propuseram uma taxonomia de tipologias de SOCs, conforme mostra a Figura 1.

Figura 1 – Taxonomia das dimensões dos SOCs.



Fonte: esquema traduzido de SOUZA; TUDHOPE; ALMEIDA (2012, p.189).

Observa-se que essa taxonomia permite perceber as dimensões das características em intrínsecas (modelo real, características próprias e essenciais de cada instrumento) e extrínsecas (criadas para um contexto de uso específico, como características acidentais). Para os autores, por meio dessas dimensões é possível determinar a classificação de um determinado tipo de SOC, o que parece ser o mais adequado para representar a distinções entre os vários tipos de SOCs (MACULAN, 2015). O SOC do tipo tesauro vem

sendo desenvolvido e empregado por diversos serviços de informação, no sentido de agregar melhorias para a recuperação da informação, possuindo uma estrutura que pode ser considerada de alta complexidade, conforme descrito a seguir.

2.1 Tesauros

A palavra “tesauro”, etimologicamente, em latim, tem a forma *thesaurus*, procedente da palavra grega *thesauros*, ficou mais conhecida em 1852, por meio do dicionário análogo “*Thesaurus of English words and phrases*” de Peter Mark Roget, publicado em Londres. Somente em 1971, com a definição fornecida nas diretrizes para construção de tesauros, apresentadas pela *United Nations Educational, Scientific and Cultural Organization* (UNESCO), ocorre uma maior “uniformidade e consistência no emprego do termo tesauro”, como um instrumento terminológico (CARVALHO, 2013, p.55).

Os tesauros são linguagens terminológicas que representam o conhecimento de determinado campo de especialidade, elaborados com base em um conjunto de regras, e são constituídos por um conjunto de termos descritores. Segundo Campos e Gomes (2006), o vocabulário de um tesauro não se constitui por palavras de linguagem natural (discurso), mas por uma lista de termos que são signos verbais que denotam um conceito em um contexto específico.

Em ambiente organizacional, Currás (1995) aponta que os tesauros têm como funções a representação dos assuntos dos documentos e o auxílio nas solicitações de buscas informacionais realizadas pelos usuários. Para a autora, na recuperação da informação o tesauro cumpre as funções de determinar os termos utilizados no sistema; definir os termos que podem ser empregados em buscas do usuário; admitir a introdução de novos termos, permitindo a constante atualização e adequação da estrutura conceitual do tesauro. Para Marroni (2006), a especificidade de um tesauro se relaciona com a sua

abrangência temática, podendo ser multidisciplinar ou de uma disciplina exclusiva.

No que se refere à estrutura de um tesauro, esta se baseia na ordenação sistemática em categorias, as quais produzem um sistema de conceitos e termos interligados por meio das relações semânticas: equivalência, hierárquica e associativa. Essa estrutura integra quatro componentes: (1) um léxico, termo que compõe uma lista de elementos descritores e não descritores; (2) uma rede paradigmática (relações definidas a priori), para indicar relações essenciais e, geralmente, estáveis entre os descritores; (4) uma rede sintagmática (*a posteriori),* para determinar as relações contingentes entre os descritores, válidas apenas no contexto particular de uso (GARDIN et al., 1968 apud CINTRA et al., 2002). Em geral, em nível conceitual, todo descritor está ligado a outro descritor (SVENONIUS, 2000), a partir de diferentes tipos de relacionamentos.

2.1.1 Relacionamentos semânticos em tesauros

O termo “relacionamentos” engloba diversas conexões entre termos, conceitos, objetos e entidades, e, nessa perspectiva, Green (2001) afirma que um

> [...] relacionamento é uma associação entre duas ou mais entidades ou entre duas ou mais classes de entidades. Para especificar um relacionamento, temos de ser capazes, em primeiro lugar, de designar todas as partes vinculadas pelo relacionamento e, em segundo lugar, de especificar a natureza dessa relação (GREEN, 2001, p.3).

Observa-se, assim, que a autora recomenda especificar a natureza da relação, identificando-se o seu significado, visando uma melhor compreensão da ligação entre os dois elementos.

A questão das relações semânticas é tratada desde muito tempo, seja por Aristóteles, que determinou quatro predicados para a determinação de relações hierárquicas, ou com Kant, que estabeleceu as relações dos

fenômenos da natureza a partir de princípios racionais e universais em situações empíricas (MACULAN, 2015). No âmbito da Ciência da Informação, há diversos autores que discorrem sobre esse tema, dentre os quais destacam-se Foskett (1973), Hutchins (1975), Dahlberg (1978; 1979), Farradane (1980), Motta (1987), Gomes (1990), Svenonius (2000), Campos (2001), Green (2001; 2008), Dodebei (2002), Cintra et al. (2002), Campos, Gomes e Motta (2004) e Kobashi (2007).

Na construção de tesauros são representados três tipos básicos de relações semânticas: de equivalência, hierárquicas e associativas, sendo que estas apresentam subdivisões de tipologias na literatura sobre o tema. Uma propriedade básica dos relacionamentos em tesauros é que eles são recíprocos, pois, para cada relação indicada entre o termo A e o termo B, tem uma relação correspondente do termo B ao termo A, sendo esta regra observada para todos os tipos de relacionamentos.

Segundo Maculan (2015), no nível da palavra (lexical), os relacionamentos semânticos lidam com equivalências: sinonímia e antonímia. Em geral, a reciprocidade desse tipo de relação deve ser expressa pelos seguintes indicadores de relacionamento: USE (termo preferencial) e UF (termo não preferencial). Estes são relacionamentos assimétricos: se TermoA USE TermoB, então, TermoB UF TermoA.

A sinonímia é a relação de equivalência semântica que existe entre duas (ou mais) palavras que têm o mesmo significado (ou quase o mesmo), dentro de um dado contexto. Sinônimos absolutos são difíceis de ocorrer, e pode-se afirmar que não há pares de palavras que tenham o mesmo significado em todos os contextos situacionais em que podem acontecer.

Maculan (2015) aponta que a antonímia é a relação de equivalência semântica que existe entre duas (ou mais) palavras que têm significados opostos que, em geral, pertencem à mesma categoria gramatical (ambos são substantivos ou adjetivos, por exemplo). Eles podem ser de três tipos básicos: complementares ou contraditórios: pares de palavras nas quais um membro possui certa propriedade semântica que o outro não possui e, em geral, são

pares de antônimos que oferecem um tipo de contraste em que não existe um meio intermediário (ex.: feminino/masculino); relacionais: são pares de palavras em que a existência de um dos termos implica a existência do outro termo (ex.: médico/paciente); graduáveis ou escalares: são pares de palavras que representam pontos finais em uma escala, podendo existir diferentes pontos intermediários entre eles (ex.: quente/frio).

Sobre os relacionamentos hierárquicos, Maculan (2015) afirma que são baseados em graus ou níveis de superordinação e subordinação, onde o termo superordenado representa uma classe ou um todo, e os termos subordinados se referem a seus membros ou partes. Em geral, a reciprocidade desse tipo de relação deve ser expressa pelos seguintes indicadores de relacionamento: BT (termo mais amplo), um descritor que é o termo superordenado; NT (termo mais específico), um descritor que é o termo subordinado, ou seja, se TermoA BT TermoB, então, TermoB NT TermoA. Todo termo subordinado deve referir-se ao mesmo tipo básico de conceito que seu termo superordenado. Assim, um termo mais amplo e o seu termo mais específico devem representar uma coisa ou uma ação ou uma propriedade, por exemplo. Em geral, as relações semânticas hierárquicas cobrem três situações logicamente diferentes: relacionamento genérico; relacionamento partitivo; relacionamento de instância.

Para Maculan (2015), os relacionamentos associativos abrangem associações entre descritores que não são equivalentes nem hierárquicos, mas que estão associados semantica e/ou conceitualmente, de tal forma que o vínculo entre eles deve ser explicitado na estrutura do tesauro, com vistas a sugerir descritores que representam conceitos adicionais para uso na indexação ou na recuperação de informações. O relacionamento associativo é simétrico e, em geral, é indicado pela abreviatura RT (termo relacionado). Essa é uma relação mais difícil de determinar, sendo importante explicitar a natureza da relação entre os descritores vinculados, evitando-se, assim, estabelecimentos inconsistentes de relação. Como uma orientação geral, haverá uma relação associativa entre dois pares de descritores se eles pertecem a quadros de referência compartilhados, sendo que um deles é parte

necessária utilizada na definição do outro (desde que não mantenham relação de equivalência e nem hierárquica). Dois descritores podem manter um relacionamento associativo sejam eles pertencentes à mesma hierarquia (em situações especiais, para orientar o usuário na localização do termo desejado) ou pertencentes a hierarquias diferentes, por diferentes motivos. O relacionamento associativo (RT) é simétrico: se TermoA RT TermoB, então, TermoB RT TermoA.

A captura de relações semânticas (abstrações de dados) é importante em aplicativos para o gerenciamento de bancos de dados, que exigem a especificação de propriedades semânticas de um relacionamento. Atualmente, a norma ISO 25964 (2011; 2013) apresenta as mais recentes recomendações para a construção de tesauros, para o seu uso na indexação, exploração e recuperação de informações em ambiente digital, orientando sobre a explicitação de relacionamentos entre termos e conceitos. Nessa perspectiva, os desenvolvedores do tesauro AGROVOC foram motivados a efetuar o refinamento dos relacionamentos de sua estrutura, conforme está apresentado na próxima seção.

3 O TESAURO AGROVOC

A *Food and Agriculture Organization* (FAO) das Nações Unidas (ONU) desenvolveu e mantém, por meio de uma comunidade de especialistas, o AGROVOC, que é um vocabulário controlado que cobre as áreas de alimentos, nutrição, agricultura, pesca, floresta, meio ambiente, entre outras. Ele é um tesauro multilíngue, publicado em 27 linguagens, e, em sua versão publicada em 15 de julho de 2016, apresenta 32.656 conceitos.

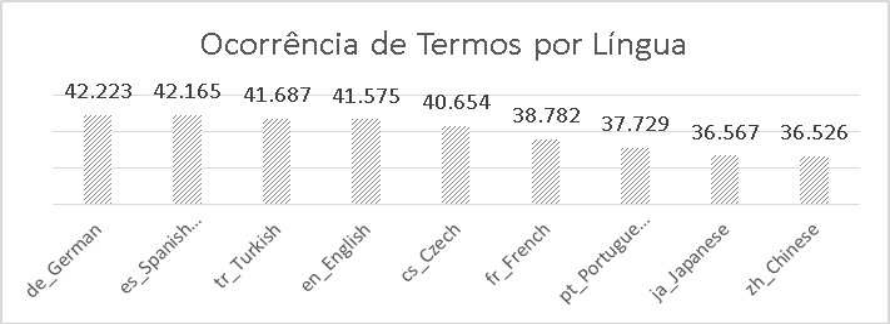
Em termos de estrutura de sistema, atualmente, o AGROVOC é um esquema de conceitos em SKOS-XL e um conjunto de *Linked Open Data* (LOD) alinhado com outros 16 sistemas multilíngues de organização do conhecimento relacionados à agricultura. É editado por meio de uma

ferramenta de edição colaborativa, baseada em *web*, o *VocBench*, e recursos RDF-SKOS.

Entre as diversas possibilidades de uso do tesauro AGROVOC, a FAO elenca as seguintes: (a) verificar o nome comum de uma planta em uma linguagem que você não domina; (b) encontrar relacionamentos entre um produto e a cultura com que ele é produzido; (c) bibliotecas podem utilizar para indexar seus documentos; (d) um usuário pode utilizar dentro do seu sistema de gestão de conteúdo para organizar documentos ou site web; (e) como um *hub* para acessar outros vocabulários disponíveis na internet.

A evolução do Tesauro AGROVOC é resumida pela FAO em cinco principais alterações: (1) início dos anos 80: publicação impressa em papel, publicação em Inglês, Espanhol e Francês, e utilizada para a indexação de publicações em ciências e tecnologia agrícola; (2) ano 2000: publicação digital e com banco de dados relacional; (3) ano 2004: houve a sua conversão para OWL, para teste; (4) ano 2009: tornou-se um recurso em SKOS; (5) ano 2017: é um esquema conceitual SKOS-XL, com conceitos em 30 idiomas, e publicado em conjunto com o *Linked Open Data* (LOD), alinhado a 16 conjuntos de dados relacionados à área da agricultura.

Gráfico 1 – Línguas em que são apresentados os termos.



Fonte: elaborado pelos autores (2017)

Como citado anteriormente, o AGROVOC é um tesauro multilíngue e apresenta seus conceitos e termos em 27 diferentes linguagens, sendo que os termos preferidos são apresentados em 25, enquanto termos alternativos são apresentados em 26 línguas.

Vemos no gráfico 1 que, com relação às línguas presentes, o Alemão (6,53%) é a que apresenta mais termos no AGROVOC, seguido por Espanhol (6,52%), Turco (6,45%) e Inglês (6,43%).O tesauro AGROVOC é composto de 67.694 (57,40%) relações hierárquicas e 50.240 (42,60%) não hierárquicas. Com relação aos sinônimos, temos para 32.656 conceitos apenas 5.731 sinônimos (17,54%), e embora apresente três conceitos com mais de 20 sinônimos, uma ampla maioria, 3.834 (11,74%) conceitos, apresenta apenas um sinônimo.

Dentre os padrões definidos pela W3C, os seguintes foram utilizados para disponibilizar a estrutura do Tesauro AGROVOC:

- *Resource Description Framework* (RDF): modelo de dados publicado em 1997, baseado na ideia de expressões sobre recursos, na forma de *subject*, *predicate*, *object,* conhecido como triplas;
- *Web Ontology Language* (OWL): linguagem de *web* semântica projetada em 2004 para representar conhecimento complexo sobre coisas, grupos de coisas e relacionamento entre coisas;
- *Linked Open Data* (LOD): é um método de publicação de dados estruturados de forma que possam ser interligados e assim mais bem utilizados em pesquisas semânticas. Sua divulgação foi iniciada em 2007;
- SPARQL *Protocol and* RDF *Query Language* (SPARQL): uma linguagem de pesquisa de dados em formato RDF publicada em 2008;
- *Simple Knowledge Organization System* (SKOS): projetado em 2009 sobre o modelo RDF para representação de vocabulários controlados, com o objetivo de facilitar o uso destes vocabulários como *linked data*;

- *Simple Knowledge Organization System eXtension for Labels* (SKOS-XL): também projetado em 2009, provê suporte adicional ao SKOS para descrição e ligação de entidades lexicais.

Assim, o Tesauro AGROVOC, por suas características apresentadas, e por estar representado em LOD foi selecionado para essa análise quantitativa sobre o refinamento realizado em sua estrutura, com vistas a propor um modelo semiautomatizado, cujas características são apresentadas a seguir.

4 SEMIAUTOMATIZAÇÃO DE RELACIONAMENTOS EM TESAUROS

Soergel et al. (2004) abordam a questão do uso de tesauros para construção de ontologias, na qual encontramos a apresentação de *rules as you go*, uma metodologia que permite a inserção e manutenção em tesauros. A proposta de um sistema com essa abordagem é explorar padrões de relacionamentos entre tipos de entidades que possam contribuir com a automatização no processo de edição de um tesauro.

Nessa proposta, os autores sugerem a definição de uma estrutura ontológica que preencha possíveis extensões do Tesauro AGROVOC e possibilite edição assistida por computador com um sistema utilizando abordagem *rules as you go* e um editor de ontologias para tornar as informações explicitadas e, também, adicionar novas informações.

Considera-se que essa abordagem pode contribuir para a melhoria de buscas em tópicos específicos, tais como: orientações para plantio e gestão de plantio (fertilização, irrigação); orientações quanto ao manejo de pestes; controle de contaminadores na cadeia alimentar; orientações para processamento seguro de alimentos; sistemas para etiquetas automatizadas de informações nutricionais; orientações sobre alimentação saudável.

A ideia principal é formular determinadas limitações que auxiliam o editor a explicitar os relacionamentos. Em uma ontologia em que podem ocorrer mais de 100 tipos de relacionamentos, um limitador de tipos, em um caso específico, mostra ao editor apenas os tipos de relacionamentos prováveis de ocorrer. Se

o limitador apontar apenas um relacionamento existente dentre os relacionamentos, somente este será recuperado, cabendo ao editor somente confirmar sua inserção.

O Tesauro AGROVOC não apresenta todos os tipos de relacionamento e não define tipos de entidade. O quadro 1 apresenta a diferença na forma de representar os relacionamentos em um tesauro comum, no Tesauro AGROVOC, e com a abordagem dessa pesquisa.

Quadro 1: As representações de um Tesauro. Exemplo Tesauro AGROVOC.

	Subject	**Entity Type**	Predicate	**Relationship Type**	Object	**Entity Type**
Tesauro Tradicional	Cellulolytic microorganisms		narrower		Cellulose	
Tesauro AGROVOC	Cellulolytic microorganisms		actsUpon		Cellulose	
Abordagem de pesquisa	Cellulolytic microorganisms	**Microorganism**	actsUpon	**General Influences**	Cellulose	**Chemical Substance**

Fonte: elaborado pelos autores (2017).

Nota-se, em **destaque** no Quadro 1, a proposta de associação de tipos de entidade aos pares *subject-object* e tipos de relacionamento aos *predicate*, sugeridos na abordagem *rules as you go*.

5 METODOLOGIA

A presente pesquisa se baseia em uma abordagem exploratória sequencial mista. Primeiramente, foram coletados os dados no Tesauro AGROVOC, no seu formato padrão. Como resultado foi obtido um conjunto de dados apresentado na forma de triplas RDF, que é uma expressão que define o relacionamento entre os elementos. Posteriormente, esses dados foram convertidos utilizando tabelas em SQL e Excel, para possibilitar a visualização dos dados no arquivo de triplas AGROVOC.

Então, a abordagem qualitativa foi empregada, na medida em que foram realizadas análises para a classificação de tipos de entidade para o *subject* e

object e classificação de tipos relacionamentos *predicate*, que não estavam disponíveis no Tesauro AGROVOC. Assim, com os conceitos e relacionamentos classificados procedeu-se à análise quantitativa.

Tendo em vista propor uma abordagem para a semiautomatização de estabelecimento de relacionamentos entre conceitos, foram realizadas análises quantitativas e qualitativas dos tipos de entidades e tipos de relacionamentos no Tesauro AGROVOC. Dessa forma, a seguir é apresentada a análise dos dados do Tesauro AGROVOC representado em modelo SKOS/SKOS-XL, no qual foi aplicada uma classificação, baseada em ontologia, dos pares *subject – object,* em tipos de entidade, e a classificação de seus relacionamentos (*predicate*).

5.1 Procedimentos Aplicados

A análise dos elementos foi realizada tendo como base suas definições de acordo com o esquema conceitual SKOS/SKOS-XL do Tesauro AGROVOC, conforme descrito abaixo:

- *Concepts* (conceitos): são quaisquer coisas que queremos representar ou “falar sobre” no domínio. São representados por termos, podem ser considerados, também, o conjunto de todos os termos utilizados para expressá-lo em várias linguagens. O conceito é representado como um objeto skos/core#Concept;
- *Terms* (termos): são os atuais termos utilizados para dar nome a um conceito. No AGROVOC é expresso pelo significado de extensão do SKOS para *labels*, SKOS-XL. Representado por meio de predicados:
    - skos-xl#prefLabel - utilizado para termos preferidos (descritores)
    - skos-xl#altLabel - utilizado para termos não preferidos;
- *Relations* (relacionamentos) - hierárquicos e não hierárquicos
    - Hierárquico: representado em SKOS pelos predicados que correspondem aos clássicos relacionamentos de tesauro

- skos/core#broader
- skos/core#narrower

- Não hierárquico: expressa uma noção de “relacionamento” entre conceitos. Representado pelos predicados
  - skos/core#related
  - vocabulário específico de relacionamentos chamado Agrontology.
- AGROVOC permite, também, relações entre *labels*.

Primeiramente realizou-se o download das triplas na base de dados do AGROVOC, no seu formato padrão (valores separados por vírgulas e campos na sintaxe N-Triplas), utilizando a ferramenta o SPARQL *Endpoint* do repositório AGROVOC (http://bit.ly/AGROVOCRepository).

Como resultado da busca no repositório, obtivemos um conjunto de dados, apresentado na forma de triplas RDF, uma expressão que define o relacionamento entre os elementos. Como padrão, as triplas RDF apresentam a expressão *Subject* (S), *Predicate* (P) e *Object* (O).

Como exemplo selecionamos neste conjunto de dados apresentados uma tripla RDF na qual um código é definido como conceito:

Subject	Predicate	Object
<http://aims.fao.org/aos/agrovoc/c_1070>	<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>	<http://www.w3.org/2004/02/skos/core#Concept>

Os resultados obtidos no conjunto de dados do Tesauro AGROVOC são apresentados a no Quadro 2.

Quadro 2 – Visão geral do conjunto de dados.

Formatos e padrões da *web* semântica utilizados na representação[1]	LOD, OWL, RDF, SKOS, SKOS-XL
Número de triplas / expressões	6.199.982
Número de conceitos	32.656
Tipos de relacionamento	127
Tipos de Relacionamentos (**P**) genéricos	70
Tipos de Relacionamentos que não envolvem conceitos	50
Tipos de Relacionamentos em que **S** é um conceito	24
Tipos Relacionamentos em que **O** é um conceito	06
Tipos de Relacionamentos entre conceitos	57
Linguagens apresentadas	27
Termo preferido em inglês	32.474
Sinônimos em inglês	9.095
Nível de detalhamento na Hierarquia de Relacionamentos	04
Nível de detalhamento na Hierarquia de Tipos de Entidade	07

Fonte: elaborado pelos autores (2017)

A partir desse conjunto de dados os procedimentos da metodologia foram aplicados, conforme descrito na próxima subseção, juntamente com as análises realizadas.

5.3 Aplicação dos Procedimentos e Análise dos Resultados

Para esta pesquisa, selecionamos e fizemos um recorte do conjunto de dados. Foram selecionados todo *subject,* que é um conceito em evidência; o

[1] A página *web* da FAO informa que o AGROVOC é um esquema de conceitos em SKOS-XL em um conjunto de dados *Linked Open Data* (LOD).

predicate, a expressão verbal que aponta o tipo de relação semântica que ocorre; e o *object,* o conceito relacionado ao sujeito.

Observando estes elementos em mãos, constatamos que o conjunto de dados original apresenta os relacionamentos não hierárquicos apenas em uma direção *subject* – *object*. Assim, foi feita uma adequação do conjunto de dados, incorporando a ele os relacionamentos inversos *object* – *subject*, indispensável à proposta de *rules as you go*.

Outros elementos não disponíveis, mas necessários à análise, à classificação em tipos de entidade para *subject* e *object* e à classificação dos *predicate* em tipos de relacionamento, foram desenvolvidos por Soergel (2017).

A classificação em tipos de entidade dos conceitos foi elaborada por Soergel (2017) a partir da *Basic Formal Ontology (BFO*[2]*),* conforme apresentada no quadro 3.

Quadro 3 – Classificação de tipos de entidade baseado na BFO, desenvolvido por Soergel.

e_1 BFO:continuant	e_1.3.2 ..BFO:realizableEntity
e_1.1 .BFO:independentContinuant	e_1.3.2.0 ...stateCondition
e_1.1.1 ..BFO:materialEentity	e_1.3.2.1 ...policiesProceduresLawsRulesAndRegulations
e_1.1.1.1 ...BFO:object	e_1.3.2.1.1 ....policiesProcedures
e_1.1.1.1.1 ....inanimateObject	e_1.3.2.1.2 ....laws
e_1.1.1.1.1.1 .....objectIncludingEquipment	e_1.3.2.1.3 ....rulesAndRegulations
e_1.1.1.1.1.2 .....physiographicFeature	e_1.3.2.2 ...BFO:role
e_1.1.1.1.2 ....animateObjectOrganism	e_1.3.2.3 ...BFO:disposition
e_1.1.1.1.2.1 .....microorganism	e_1.3.2.3.1 ....diseaseOrDisorder
e_1.1.1.1.2.2 .....macroorganism	e_1.3.2.3.2....BFO:function
e_1.1.1.1.3 ....bodyPart	e_1.3.2.3.2.1.....functionOfMaterialOrChemicalSubstances
e_1.1.1.2 ...material	e_1.3.2.3.2.1.....functionOfMaterialOrChemicalSubstances
e_1.1.1.2.1 ....otherMaterial	e_1.3.2.3.2.2 .....biologicFunction
e_1.1.1.2.2 ....chemicalSubstance	e_2 BFO:occurrent
e_1.1.1.2.3 ....foodProduct	e_2.1 .BFO:processBroad
e_1.1.1.4 ...BFO:fiatObjectPart	e_2.1.1 ..process
e_1.1.1.5 ...BFO:object aggregate	e_2.1.1.1 ...processHappening
e_1.1.1.5.1 ....organization	e_2.1.1.2 ...OBI:plannedProcessOrActivity
e_1.1.1.5.2 ....population	e_2.1.1.2.1 ....activity
	e_2.1.1.2.2 ....methodTechnique

[2] *Basic Formal Ontology* (BFO) é uma pequena e genuína ontologia de nível superior desenvolvida para dar suporte a recuperação, análise e integração de informação no domínio científico e outros domínios, utilizada (não se refere à antologia), por mais de 250 empreendimentos orientados a ontologia. Desenvolvido por Barry Smith e Pierre Grenon, disponível em http://bit.ly/BFOntology.

e_1.1.2 ..BFO:immaterialEntity	e_2.1.1.2.2.1 .....materialMethodTechnique
e_1.1.2.1 ...BFO:continuantFiatBoundary	e_2.1.1.2.2.1.1 .....materialMethodTechniqueDoing
e_1.1.2.2 ...BFO:site	e_2.1.1.2.2.1.2 .....materialMethodTechniqueResearch
e_1.1.2.3 ...BFO:spatialRegion	e_2.1.1.2.2.2 .....statisticalMethodTechnique
e_1.2 .BFO:genericallyDependentCcontinuant	e_2.1.2 ..behavior
e_1.2 .BFO:genericallyDependentContinuant	e_2.1.3 ..service
e_1.2.1 ..informationArtifact	e_2.1.4 ..eventType
e_1.3 .BFO:specifically dependent continuant	e_2.1.5 ..BFO:history
e_1.3 .BFO:specificallyDependentContinuant	e_2.2 .BFO:processBoundary
e_1.3.1 ..BFO: quality == property	e_2.3 .BFO:spatiotemporalRegion
e_1.3.1.1 ...measureMetric	e_2.4 .BFO:temporalRregion
e_1.3.1.2 ...physicalState	e_2.4.1 ..BFO:zeroDimensionalTemporalRegion
e_1.3.1.3 ...soilType	e_2.4.2 ..BFO:oneDimensionalTemporalRegion
e_1.3.1.4 ...geographicAreaType	e_2.5 .timeRelatedConcepts
e_1.3.1.5 ...climate	e_3 nonBFOEntities
e_1.3.1.6 ...buildingType	e_3.1 .namedPlaceOrLocation
e_1.3.1.7 ...taxonLevel	e_3.3 .scientificScholarlyArea
e_1.3.1.8 ...taxonProperty	e_3.4 .workersProfessions
e_1.3.1.9 ...organizationType	e_3.5 .economicSector
e_1.3.1.A ...populationType	e_3.5 .economic Sector
e_1.3.1.B ...developmental StageAgeGroup	e_3.6 .agricultureSystem
e_1.3.1.B ...organismProperty	e_3.7 .socioEconomicSystem
e_1.3.1.B ...developmentalStageAgeGroup	e_3.8 .requirement
e_1.3.1.C ...BFO:relational quality	e_3.9 .offspringRelation
	e_3.A .ration
	e_3.B .resource
	e_3.C .lossDamage
	e_3.X .MiscellaneousBiologic
	e_3.Y .MiscellaneousFinanceRelated
	e_3.Z .MiscellaneousOther

Fonte: elaborado pelos autores (2017)

Com esta classificação criou-se uma hierarquia de tipos de entidade completa da estrutura do Tesauro AGROVOC, iniciando do geral para o específico. A análise desta hierarquia demonstrou a existência de grandes segmentos em que diversos conceitos foram atribuídos a um mesmo tipo de entidade. Como exemplo da entidade organismos, que abrange os tipos *Animate Object Organism*, *Microorganism* e *Macroorganism*.

Percebe-se, assim, que apesar do Tesauro AGROVOC possuir um grande número de conceitos, mais da metade, 20.293, são referentes a entidade "*organism*" e estão distribuídos em 18.423 "*macroorganism*", 1.651 "*microorganism*" e 219 "*animate object organism*". Observamos, também, que a entidade "*material*" possui 2.950 conceitos relacionados a ela e que há um pequeno agrupamento, de somente 13 conceitos, na entidade "*population*". Verifica-se, assim, que o escopo do Tesauro AGROVOC não é amplo suficiente na cobertura dos conceitos da área.

Conforme citado anteriormente, e apresentado no quadro 01, para os procedimentos de análise foi gerada uma tabela em Excel, a partir da qual foram gerados os resultados. O conteúdo foi preenchido a partir do recorte dos relacionamentos entre conceitos e foi estruturado da seguinte forma:

> subjectCode; subjectLabel; sEntityTypeCode; sEntityTypeLabel; predicateRelationshipTypeCode; predicateLabel; objectCode; objectLabel; oEntityTypeCode; oEntityTypeLabel

A partir dessa tabela, foram geradas fórmulas em Excel para análise e considerações alçadas.

A análise dos relacionamentos semânticos entre conceitos apresentou 57 tipos de relacionamentos em uso no Tesauro AGROVOC. Estes se desdobram em 117.934, conforme Quadro 4.

Quadro 4 – Frequência de tipos de relacionamento em AGROVOC por instância. Todos os relacionamentos.

r_03	**Taxonomic relationship**	**7.534**	**r_03i**	**Taxonomic relationship**	**7.534**
r_04	**$Spatial relations**	**565**	**r_04i**	**Spatial relations**	**565**
r_05	**Temporal relations**	**49**	**r_05i**	**Temporal relations**	**49**
r_06	**Quantitative relationship**	**60**	**r_06i**	**Quantitative relationship**	**60**
r_07	**Includes**	**2.932**	**r_07i**	**Included in**	**2.932**
r_08	**Has part**	**1.168**	**r_08i**	**Is part of**	**1.168**
r_09	**General related**	**714**	**r_09i**	**General related**	**714**
r_10	**General affects**	**187**	**r_10i**	**General affects**	**187**
r_11	**General causes**	**530**	**r_11i**	**General causes**	**530**
r_12	**General influences**	**4.068**	**r_12i**	**General influences**	**4.068**
r_13	**Instrumental relations**	**4.060**	**r_13i**	**Instrumental relations**	**4.060**
r_14	**Production relationship**	**955**	**r_14i**	**Production relationship**	**955**

r_15	**Narrower**	**33.847**	**r_15i**	**Broader**	**33.847**
r_16	**Related**	**2.290**	**r_16i**	**Related**	**2.290**
Relationship subtotal 58.967			**Inverse Relationship subtotal 58.967**		
Total Relationships 117.934					

Fonte: elaborado pelos autores (2017)

Observa-se que dentre estes 117.934 relacionamentos diretos e inversos existem 67.694 relacionamentos hierárquicos (*Narrower* e *Broader*) e 50.240 relacionamentos não-hierárquicos (demais relacionamentos).

A partir da estrutura de dados gerada foi possível fazer a junção das classificações em tipos de entidades e tipos de relações semânticas. Obtivemos um panorama em que se observou a predominância de "*organism*", assim fizemos um levantamento de ocorrências de relacionamentos considerando apenas "organism" e outro considerando os não "organism".

Quadro 5 – Frequência de tipos de relacionamento em AGROVOC por instância.Instâncias de Relacionamento onde o conceito *subject* ***pertence*** ao tipo entidade ***Organism***.

r_03	**Taxonomic relationship**	**7.529**	**r_03i**	**Taxonomic relationship**	**00**
r_04	**$Spatial relations**	**00**	**r_04i**	**Spatial relations**	**00**
r_05	**Temporal relations**	**00**	**r_05i**	**Temporal relations**	**01**
r_06	**Quantitative relationship**	**01**	**r_06i**	**Quantitative relationship**	**01**
r_07	**Includes**	**565**	**r_07i**	**Included in**	**786**
r_08	**Has part**	**22**	**r_08i**	**Is part of**	**35**
r_09	**General related**	**22**	**r_09i**	**General related**	**40**
r_10	**General affects**	**03**	**r_10i**	**General affects**	**15**
r_11	**General causes**	**213**	**r_11i**	**General causes**	**03**
r_12	**General influences**	**1.379**	**r_12i**	**General influences**	**1.427**
r_13	**Instrumental relations**	**1.501**	**r_13i**	**Instrumental relations**	**319**
r_14	**Production relationship**	**435**	**r_14i**	**Production relationship**	**18**
r_15	**Narrower**	**20.661**	**r_15i**	**Broader**	**20.613**
r_16	**Related**	**1.232**	**r_16i**	**Related**	**1.232**
Relationship subtotal 33. 563			**Inverse Relationship subtotal 24. 490**		
Total Relationships 58. 053					

Fonte: elaborado pelos autores (2017)

Quadro 6 – Frequência de tipos de relacionamento em AGROVOC por instância. Instâncias de Relacionamento onde o conceito *subject* ***não pertence*** ao tipo entidade ***Organism***.

r_03	**Taxonomic relationship**	**05**	**r_03i**	**Taxonomic relationship**	**7.534**
r_04	**$Spatial relations**	**565**	**r_04i**	**Spatial relations**	**565**
r_05	**Temporal relations**	**49**	**r_05i**	**Temporal relations**	**48**
r_06	**Quantitative relationship**	**59**	**r_06i**	**Quantitative relationship**	**59**
r_07	**Includes**	**2.367**	**r_07i**	**Included in**	**2.146**
r_08	**Has part**	**1.146**	**r_08i**	**Is part of**	**1.133**
r_09	**General related**	**692**	**r_09i**	**General related**	**674**
r_10	**General affects**	**184**	**r_10i**	**General affects**	**172**
r_11	**General causes**	**317**	**r_11i**	**General causes**	**527**
r_12	**General influences**	**2.689**	**r_12i**	**General influences**	**2.641**
r_13	**Instrumental relations**	**2.559**	**r_13i**	**Instrumental relations**	**3.741**
r_14	**Production relationship**	**520**	**r_14i**	**Production relationship**	**937**
r_15	**Narrower**	**13.186**	**r_15i**	**Broader**	**13.234**
r_16	**Related**	**1.058**	**r_16i**	**Related**	**1.058**
Relationship subtotal25.404			**Inverse Relationship subtotal**		**34.477**
Total Relationships 59.881					

Fonte: elaborado pelos autores (2017)

A junção das classificações possibilitou observar a predominância de instâncias de relacionamento para a entidade do tipo organismo, conforme demonstrado no Quadro 4.

Para a uma análise mais específica de como ocorrem os relacionamentos no Tesauro AGROVOC, o Quadro 7 apresenta um exemplo:

Quadro 7 – Exemplo de ocorrência de relações no AGROVOC

Relationship type: spatiallyIncludes		
Qualidade	**(Subject , Object)**	**Ocorrências**
bom	(*namedPlaceOrLocation , namedPlaceOrLocation*) E.g. (*Canadá , Fraser river*)	537
ruim	(*physiographicFeature , physiographicFeature*) E.g. (*lowland , valleys*)	3
ruim	(*physiographicFeature , namedPlaceOrLocation*) E.g. (*Boreal forests , Arctic tundra*)	2

Fonte: elaborado pelos autores (2017)

Neste exemplo, pode-se verificar, no caso de *spatiallyIncludes*, que há uma predominância de relacionamento entre *namedPlaceOrLocation* (nomeOuLugar) o que está correto, pois em sua definição *spatiallyIncludes* diz que Y é uma parte inerente ou inalienável de X, sendo que Y é aquilo que está intimamente unido a X, pois diz respeito ao próprio ser X. Ou seja, Y é uma parte de X (animado ou inanimado) e que é inseparável ou intrínseco de X por natureza. Definimos então como um bom relacionamento (*namedPlaceOrLocation , namedPlaceOrLocation*) determinado pela definição do tipo de relacionamento. Enquanto que (*physiographicFeature , physiographicFeature*) e (*physiographicFeature , namedPlaceOrLocation*) são relacionamentos ruins para a elaboração de regras.

Por meio destas análises preliminares verificamos que, embora tenha definido os tipos de relacionamento a serem utilizados, claramente nem todos os relacionamentos são utilizados ou são pouco utilizados. Isto significa que o refinamento do AGROVOC não está completo.

6 CONSIDERAÇÕES FINAIS

A proposta desta pesquisa se respalda em um dos focos de pesquisa na CI, a organização do conhecimento, que, como definido por Barité et al. (2013), é o estudo das leis, princípios e procedimentos pelos quais se estrutura o conhecimento especializado em qualquer disciplina. Assim, tem por finalidade representar tematicamente esse conhecimento para viabilizar a recuperação da informação contida em recursos informacionais de forma eficiente e rápida, atendendo às necessidades dos usuários. Nesse sentido, e com o intuito de facilitar a recuperação da informação, busca-se desenvolver metodologias que possam contribuir de forma sistemática com a elaboração, desenvolvimento e aplicação de ferramentas para esse fim.

A análise qualitativa, feita na classificação dos conceitos em tipos de entidade e na classificação hierárquica dos relacionamentos, desenvolvida por Soergel (2017), juntamente com esta análise quantitativa servem de insumo

para a pesquisa mais ampla que busca reconhecer padrões de relacionamento entre tipos de entidade específicos que permitam a elaboração de um sistema com base na criação de regras, conforme é feita a edição do tesauro (*rules as you go*).

Estes resultados apontam para a entidade *organism* como recorte possível para o desenvolvimento da tese que propõe um modelo de sistema de refinamento de tesauro com base na abordagem *rules as you go*.

REFERÊNCIAS

BARITÉ, M. et al. **Diccionario de Organización del Conocimiento**: clasificación, indización, terminología. 5a. ed. Montevideo: PRODIC, 2013.

BOCCATO, V. R. C.; BISCALCHIN, R. As dimensões culturais no contexto da construção de vocabulários controlados multilíngues. **Rev. Interam. Bibliot**, Medellín, v. 37, n. 3, p. 237-250, Dec. 2014.

BRÄSCHER, M.; CAFÉ, L.. Organização da Informação ou Organização do Conhecimento? In: SMIT, J. W.; GINEZ DE LARA, M. L. (Org.). **Temas de pesquisa em Ciência da Informação no Brasil**. São Paulo: Escola de Comunicações e Artes/USP, 2010. p. 8-103. Disponível em: <http://www3.eca.usp.br/sites/default/files/form/ata/pos/ppgci/publicacoes%20-%20temasdepesquisas.pdf>. Acesso em: 20 jan 2017.

CAFÉ, L. M. A.; BARROS, C. M.; SANTOS, V. C. O conceito de Organização do Conhecimento nas revistas brasileiras de Ciência da Informação. **Rev. Interam. Bibliot**, Medellín, v. 37, n. 3, p. 201-214, Dec. 2014.

CAFÉ, L.; BRÄSCHER, M. Organização do Conhecimento: teorias semânticas como base para estudo e representação de conceitos. **Inf. & Inf.**, Londrina, v. 16, n. 2, p. 25-51, dez. 2011. Disponível em: <http://www.uel.br/revistas/uel/index.php/informacao/article/view/10388/9282>. Acesso em: 25 fev. 2017.

CAMPOS, M. L. A. **Linguagem documentária:** teorias que fundamentam sua elaboração. Rio de Janeiro: EUFF, 2001.

CAMPOS, M. L. A.; GOMES, H. E. Metodologia de elaboração de tesauro conceitual: a categorização como princípio norteador. **Perspect. Ciênc. Inf.**, Belo Horizonte, v. 11, n. 3, p. 348-359, set./dez. 2006. Disponível em: <http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/view/273/66>. Acesso em: 05 março 2016.

CAMPOS, M. L. A.; GOMES, H. E.; MOTTA, D. F. **Manual de elaboração de tesauro.** Rio de Janeiro: BITI, 2004.

CARVALHO, S. A. L. **Terminologia e Documentação: um estudo terminográfico sobre performance musical**. 2013. 188f. Dissertação (Mestrado em Ciência da Informação) – Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2013.

CINTRA, A. M. M. et al. **Para entender as linguagens documentárias**. 2. ed. rev. ampl. São Paulo: Polis, 2002.

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

DAHLBERG, I. A referent-oriented, analytical concept theory of Interconcept. **International Classification**, v. 5, n. 3, p. 122-151, 1978.

DAHLBERG, I. Teoria da classificação, ontem e hoje. Tradução de Henry B. Cox. In: CONFERÊNCIA BRASILEIRA DE CLASSIFICAÇÃO BIBLIOGRÁFICA, 12-17 de setembro de 1972, Rio de Janeiro. **Anais...** Brasília, IBICT/ABDF, v. 1, p. 352-370, 1979.

DODEBEI, V. L. D. **Tesauro:** linguagem de representação da memória documentária. Niterói: Intertexto, 2002.

FARRADANE, J. Relational Indexing: part I e part ll. **Journal of Information Science**, n.1, p. 267-276; 313-324, 1980.

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

GOMES, H. E. (Org.). **Manual de elaboração de tesauros monolíngues**. Brasília: Programa Nacional de Bibliotecas de Instituições de Ensino Superior, 1990.

GREEN, R. Overview of relationship in knowledge organization. In: BEAN, C. A.; GREEN, R. (Ed.). **Relationship in knowledge organization**. Dordrecht: Kluwer, 2001. Chapter 1, p. 3-18.

GREEN, R. Relationships in knowledge organization. **Knowledge Organization**, v. 35, n. 2-3, p. 150-159, 2008.

HODGE, G. **Systems of knowledge organization for digital libraries**: beyond traditional authorities files. Washington, DC: Council on Library and Information Resources, 2000. Disponível em: <https://www.clir.org/pubs/reports/pub91/pub91.pdf>. Acesso em: 5 out. 2016.

HUTCHINS, W. J. **Languages of indexing and classification:** a linguistic study of structures and functions. Stevenage: Herts Peter Peregrinus, 1975. (Librarianship and information studies, 3).

INTERNATIONAL STANDARD ORGANIZATION. **ISO 25964**: thesauri and interoperability with other vocabularies. Part 1: thesauri for information retrieval. Geneve: International Standard Organization, 2011.

INTERNATIONAL STANDARD ORGANIZATION. **ISO 25964**: thesauri and interoperability with other vocabularies. Part 2: interoperability with other vocabularies. Geneve: International Standard Organization, 2013.

KOBASHI, N. Y. Fundamentos semânticos e pragmáticos da construção de instrumentos de representação de informação. **DataGramaZero - Revista de Ciência da Informação**, v. 8, n. 6, dez. 2007.

MACULAN, B. C. M. S. **Estudo e aplicação de metodologia para reengenharia de tesauro**: remodelagem do THESAGRO. 2015. 339f. Tese (Doutorado em Ciência da Informação) - Escola de Ciência da Informação da Universidade Federal de Minas Gerais, Belo Horizonte, 2015.

MARRONI, G. N. B. **Identificação e delimitação de relações associativas em tesauros**: um estudo de caso na área do direito do trabalho. 2006. 127f. Dissertação (Mestrado em Ciência da Informação) – Universidade de Brasília, Brasília, 2006.

MOTTA, D. F. **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. **J. Digital Inf.**, v.4, n.4, p. 1-23, 2004. Disponível em: <https://journals.tdl.org/jodi/index.php/jodi/article/view/112/111>. Acesso em: 05 março 2016.

SOERGEL, D. **Entity type assignments.** [mensagem pessoal]. Mensagem recebida por <deciowbj@gmail.com>. Acesso em: 18 fev. 2017.

SOUZA, R. R.; TUDHOPE, D.; ALMEIDA, M. B. **Towards a taxonomy of KOS:** dimensions for classifying knowledge organization systems. **Knowledge Organization**, v. 39, n. 3, p. 179-192, 2012. Disponível em: <http://mba.eci.ufmg.br/ downloads/ Souza_Tudhope_Almeida_-_KOS_Taxonomy.Submitted.pdf>. Acesso em: 10 out. 2016.

SVENONIUS, E. **The intellectual foundations of information organization**. Cambridge: The MIT Press, 2000.

ZENG, M. L. Knowledge organization systems (KOS). **Knowledge Organization**, Frankfurt, v. 35, n. 2-3, p. 160-182, Jan. 2008.

Title

Relationsip semi-automation in thesauri: a proposal for semantic relationship refinement based on thesaurus agrovoc.

ABSTRACT
Introduction: Thesauri are tools that contribute to information retrieval in information services like digital data bases and digital libraries. **Objective:** This study aims to present a quantitative analysis on the refined sematic structure of the AGROVOC Thesaurus and propose a semi-automated model for semantic relationship refinement of thesauri. **Methodology:** To carry out this study, we used data from the AGROVOC Thesaurus represented in a refined SKOS-XL model. This data has been qualitatively aggregated in the classification of concepts by entity types as well as the hierarchical classification of relationships done by Soergel. From this base, we conducted a quantitative analysis of the relationship types. **Results:** Results from the quantitative data analysis showed that the refinement of the AGROVOC Thesaurus is not yet complete. Most Related Term relationships seem to have been refined, but hierarchical relationships (Broader / Narrower) have not. **Conclusion:** In sum, this study demonstrates that quantitative analysis can shed light on the structure of a thesaurus and indicate areas where improvements are possible.

Keywords: Knowledge Organization System. Thesaurus. Sematic relations. AGROVOC.

Titulo

Semi-automatización de relaciones en tesauros: una propuesta para el refinamiento de relaciones semánticas a partir del tesauro agrovoc

Resumen:
Introducción: Los tesauros son herramientas que contribuyen para la recuperación de la información en servicios, tales como bases de datos y bibliotecas digitales. **Objetivo:** Presentar un análisis cuantitativo sobre el refinamiento de la estructura semántica del Tesauro AGROVOC, con el fin de proponer un modelo parcialmente

automatizado para el refinamiento de las relaciones semánticas en los tesauros. **Metodología:** se utilizaron los datos del Tesauro AGROVOC, refinado y representado en el modelo SKOS-XL, agregados al análisis cualitativo de la clasificación de conceptos de tipos de entidades y de la clasificación jerárquica de relaciones propuesta por Soergel. Con dicha base se realizó un análisis cuantitativo de las clases de relaciones. **Resultados:** El resultado del análisis cuantitativo demuestra que el refinamiento del AGROVOC está incompleto. La mayoría de las relaciones, *related term* parecen estar refinadas, aunque las jerárquicas (*broader/narrower*) parecen no estarlo. **Conclusiones:** Los resultados demuestran que el análisis cuantitativo trae esclarecimientos sobre la estructura del tesauro indicando áreas en las que es posible implantar mejorías.

Palabras clave: Sistema de Organización del Conocimiento. Tesauro. Relaciones semánticas. AGROVOC.

Recebido em: 20.06.2017
Aceito em: 10.12.2017