XXI ENANCIB
Encontro Nacional de Pesquisa em Ciência da Informação
50 anos de Ciência da Informação no Brasil: diversidade, saberes e transformação social
Rio de Janeiro • 25 a 29 de outubro de 2021

XXI Encontro Nacional de Pesquisa em Ciência da Informação – XXI ENANCIB

GT-8 – Informação e Tecnologia

A CONSTRUÇÃO DE AMBIENTES SEMÂNTICOS PARA A RECUPERAÇÃO INTELIGENTE DA INFORMAÇÃO: uma prova de conceito

THE CONSTRUCTION OF SEMANTIC ENVIRONMENTS FOR INTELLIGENT RETRIEVAL OF INFORMATION: a proof of concept

Nilson Theobald Barbosa – UFF/UFRJ
Maria Luiza de Almeida Campos – UFF/UFBA

Modalidade: Trabalho Completo

Resumo: Os recentes avanços nas tecnologias e na sua capacidade de gerar, armazenar e compartilhar informações não vêm sendo acompanhados pela nossa capacidade de recuperar de forma inteligente toda esta informação disponível. O presente artigo visa apresentar parte de nossa pesquisa de doutorado com uma abordagem para a compatibilização de vocabulários heterogêneos com vistas a permitir a recuperação da informação em diferentes bases de dados. Temos como pressuposto a garantia que sejam mantidos sem alteração os vocabulários originais, uma vez que devido ao grande número de bases e vocabulários legados e diferentes tecnologias sendo usadas para criação de novas bases e vocabulários não será possível que este trabalho de compatibilização seja realizado de forma manual e mediada. Com relação às bases metodológicas utilizadas neste artigo, nossa pesquisa pode ser definida como básica e comparativa, visando a criação de uma prova de conceito. Para a realização desta prova de conceito utilizamos um recorte amostral extraído de três vocabulários utilizados pela Embrapa, com vistas a discutir e propor procedimentos que pudessem de forma automatizada criar um espaço semântico voltado para a recuperação da informação.

Palavras-Chave: Interoperabilidade; Sistemas de organização do conhecimento; Recuperação da informação.

Abstract: *Recent advances in technologies and their ability to generate, store and share information have not been accompanied by our ability to intelligently retrieve all this available information. This article aims to present part of our research with an approach to the compatibilization of heterogeneous vocabularies with a view to allowing the retrieval of information in different databases. We are guaranteed to maintain without change the original vocabularies, since due to the large number of legacy bases and vocabularies and different technologies being used to create new bases and vocabularies it will not be possible for this work of compatibilization to be carried out manually and mediated. Regarding the*

methodological bases used in this article, our research can be defined as basic and comparative, aiming at the creation of a proof of concept. To perform this proof of concept we used a sample cut out of three vocabularies used by Embrapa, aiming to discuss and propose procedures that could in an automated way create a semantic space aimed at information retrieval.

Keywords: *Interoperability; Knowledge organization systems; Information retrieval.*

1 INTRODUÇÃO

Temos hoje um grande avanço tecnológico e computacional que contribui com a geração contínua de um grande volume de dados digitais, que cria redes com grande desempenho, sejam as redes de fibra ótica, as redes sem fio ou a vindoura rede 5G, e que usa repositórios “infinitos” e fornece computação de alto poder de desempenho na palma da mão. Apesar disso, temos um limitador de forte impacto para a utilização plena de toda esta possível informação. Uma miríade de repositórios de dados e seus conteúdos com todos os tipos de dados se multiplicam exponencialmente, sendo que, a cada período considerado, o crescimento continua e de forma mais acentuada que no período anterior. Estes repositórios, seus dados e suas linguagens de indexação heterogêneas em todos os seus níveis, seja dentro das organizações seja na Web como um todo, ainda não são capazes de oferecer aos seus usuários uma recuperação plena e semântica da informação que está disponível.

Desde a invenção do alfabeto, como um processo que possibilita a gravação e o armazenamento de informações, passando por Gutenberg, num processo de disseminação da informação, e até a revolução dos computadores, que podemos chamar de um processo de manipulação da informação, cada uma destas etapas foi capaz de gerar novas oportunidades e novos desafios, e as tecnologias digitais do presente não são uma exceção. Se, por um lado oferecem novas formas de disponibilidade, acessibilidade e utilização, por outro lado, esta multiplicidade de recursos trazem uma grande variedade de formas de representação e indexação que dificultam a recuperação semântica desta informação entre diferentes bases de dados organizadas por diferentes sistemas de organização do conhecimento (SOCs).

Os problemas que enfrentamos ao buscar a interoperabilidade entre diferentes sistemas de organização do conhecimento se apresentam muito próximos tanto quando consideramos ambientes legados e tanto quando consideramos ambientes construídos com as novas tecnologias voltadas para uma web semântica. A interoperabilidade de redes entre

computadores e entre bases de dados é uma questão já bem resolvida pela tecnologia, a compatibilização entre termos de mesma grafia em diferentes vocabulários já é bem realizada pelo alto desempenho computacional disponível e facilitada pela conectividade, mas transpor a barreira semântica, em que precisamos compatibilizar conceitos, tenham eles a mesma expressão verbal ou não, é uma questão ainda a ser resolvida.

Portanto, temos hoje este problema, ou seja, como tornar possível um processo de recuperação inteligente de informações com tamanha diversidade e volume de geração de dados. Esta diversidade advém do espaço composto por diferentes áreas do conhecimento, diferentes ontologias, diferentes vocabulários, diferentes línguas e diferentes culturas. Mesmo em ambientes controlados, como uma única empresa, com um pequeno número de repositórios de dados ou um único domínio, por exemplo, temos dificuldades em estabelecer meios unificados de recuperação de informações que atendam a todos os produtores e consumidores da informação.

Foi com esta perspectiva que escolhemos para compor o campo empírico de nosso trabalho o ambiente informacional da Empresa Brasileira de Agropecuária – EMBRAPA. Temos hoje na EMBRAPA uma reunião de diversos fatores a serem combinados, de forma condizente com a diversidade de áreas do conhecimento abordadas pela empresa e a diversidade de grupos de pesquisa, e também com o tamanho da empresa, seus processos de gestão e sua necessidade de comunicação nacional e internacional.

Todo este ambiente forma um espaço conceitual fortemente heterogêneo em sua essência, com grupos de pesquisadores atuando de forma interdisciplinar, e consequentemente utilizando diferentes recursos informacionais, de diferentes formas e tecnologias, fazendo que o desafio a ser enfrentado passe a ser conseguir buscas que visem recuperar informações de todos estes recursos informacionais, garantindo que as idiossincrasias locais sejam respeitadas, ou seja, respeitando a origem semântica de cada comunidade envolvida, mas permitindo o acesso a diferentes públicos. Isto certamente passa pela possibilidade de comunicação dos vocabulários utilizados para indexação dos recursos informacionais disponíveis.

Apesar de parecer tentador criar sistemas de indexação e vocabulários unificados para resolver o problema da necessidade de compatibilização de sistemas de organização do conhecimento, em um ambiente aberto e não controlado nem sempre (e quase nunca) é possível recorrer a esta solução, e precisamos partir para soluções que possibilitem

recuperar informações de bases indexadas por sistemas heterogêneos sem fazer alterações nestas bases ou em seus vocabulários através do estabelecimento de correspondências e mapeamentos de conceitos, e não simplesmente de termos verbais, entre estes vocabulários.

É cada vez mais claro que, por um lado, a exploração deste aspecto do desenvolvimento humano passa inexoravelmente pela capacidade de acesso e cognição da informação disponível e as ciências humanas têm hoje ao seu dispor novos métodos de colaboração que permitem às redes internacionais criarem e explorarem imensas bases de dados que se renovam e se multiplicam em tempo real. Mas, por outro lado, esta profunda mudança epistemológica das ciências do homem somente será completada quando for possível caminhar para a adoção de uma linguagem de descrição comum que seja equivalente a dos elementos para a química ou dos meridianos e paralelos para a geografia. Como diz Lévy, adotando uma metalinguagem como esta, "as ciências do homem passariam de um estado taciturno e fragmentado a um estado onde a explicitação e a interconexão semântica das ideias e dos dados se tornariam a nova moeda corrente da prática científica" (LÉVY, 2014, p. 196).

Portanto, este é o problema a ser abordado neste trabalho, ou seja, a dificuldade de recuperar informações distribuídas em ambientes abertos heterogêneos, em diferentes bases de dados, com diferentes formatos e organizados em diferentes sistemas de organização do conhecimento.

Apresentamos como solução a possibilidade de compatibilização semântica automatizada de diferentes sistemas de organização do conhecimento sem a alteração de seus ambientes originais, com a utilização de metalinguagens, que parecem ser capazes de fornecer as bases teóricas para o desempenho desta tarefa. Estas metalinguagens devem ser formadas como uma linguagem intermediária entre os diferentes vocabulários fonte possibilitando diferentes atores navegarem por esta linguagem e, de forma contextual e semântica, recuperarem a informação pretendida, não mais com base em simples comparações de cadeias de caracteres, mas sim em seu significado.

Neste artigo, em que apresentamos parte dos estudos e resultados obtidos em nossa tese de doutorado defendida no PPGCI/UFF, esta questão será endereçada a partir desta introdução e dos procedimentos metodológicos colocados na seção 2, onde nos apoiamos para o desenvolvimento do trabalho. Na seção 3 apresentamos uma discussão sobre as

linguagens intermediárias e as coordenadas semânticas, como um suporte teórico, para que na seção 4 seja apresentado uma proposta de caminho para a criação de espaços semânticos que propiciem um processo de recuperação inteligente da informação. Alguns dos resultados obtidos em nosso trabalho de tese, relativos à construção de um espaço semânticos para recuperação da informação, são mostrados na seção 5. Finalmente, na seção 6 apresentamos nossas considerações finais.

2 METODOLOGIA

No escopo deste artigo, quanto à sua natureza, apontamos esta ser uma pesquisa básica, conforme Gerhardt e Silveira (2009, p.34), que apontam objetivar a geração de novos conhecimentos sem necessariamente uma aplicação prática prevista e que envolve verdades e interesses universais. Conforme Minayo (2010), a pesquisa básica não tem objeto prático em seu projeto inicial, mas preocupa-se com o avanço do conhecimento por meio da construção de teorias e teste delas. Em nossa pesquisa, não estabelecemos uma aplicação prática imediata, mas avançamos na efetivação de uma prova de conceito, que é uma parte do desenvolvimento de um protótipo e de uma aplicação prática completa e que pode ser considerada um teste das teorias expostas.

Como procedimento metodológico para análise dos métodos estudados, apresentamos a utilização do método comparativo. Conforme Gil (2008), este método se desenvolve a partir da investigação de indivíduos, classes, fenômenos, técnicas ou fatos, visando ressaltar as diferenças e semelhanças entre eles e, para Lakatos e Marconi (2007), permite analisar o dado concreto, deduzindo do mesmo o que se apresenta como um elemento constante, abstrato ou geral.

O foco do campo empírico de nossa pesquisa foi a utilização de três vocabulários utilizados pela Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA) e largamente difundidos no país e no mundo, voltados para indexação informações na área de agropecuária. Estes vocabulários são o AGROVOC Multilingual Thesaurus (AGROVOC, 2021), o THESAGRO – Thesaurus Agrícola Nacional (THESAGRO, 2020) e o CAB Thesaurus (CABI, 2019), dos quais foram extraídos os termos que participaram do experimento, um total de 100 termos relativos ao campo da Bioeconomia, a partir de uma relação definidas pela EMBRAPA como termos representativos desta área de estudos. A distribuição dos termos entre os vocabulários ocorreu conforme o quadro 1.

Quadro 1 – Total de termos selecionados em cada SOC

Ocorrência	Total de termos
Agrovoc-Thesagro-CAB	33
Agrovoc-Thesagro	16
Agrovoc-CAB	8
Thesagro-CAB	7
Agrovoc	14
Thesagro	4
CAB	18

Fonte: Elaborado pelo autor

Os procedimentos adotados sobre a amostra escolhida estão explicitados na seção 5 deste artigo, a saber, a elaboração do registro do conceito para cada termo, e os mapeamentos e técnicas algorítmicas utilizadas.

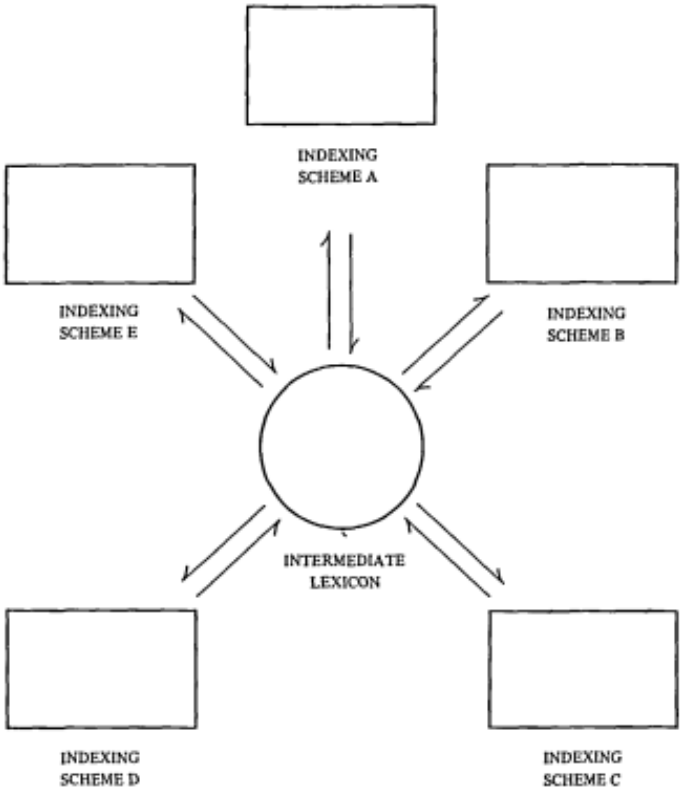
3 AS LINGUAGENS INTERMEDIÁRIAS E AS COORDENADAS SEMÂNTICAS

Neste trabalho nosso interesse se volta para uma abordagem de criação de linguagens intermediárias como uma estratégia para a compatibilização vocabulários, por considerar que as necessidades de compatibilização de linguagens para o momento atual, com vistas a uma Web Semântica devem ter por base este arcabouço teórico da Ciência da Informação. Consideramos, neste estudo, portanto, que estes métodos da Ciência da Informação têm uma forte contribuição a dar atualmente e colocamos um detalhamento sobre este assunto.

Horsnell (1975) já mostrava um método em que os índices existentes entre diferentes centros possam ser compatíveis, seja (i) promovendo a reconciliação de tesauros, onde termos de cada um de vários tesauros independentes são reconciliados resultando em uma lista mestra ou um grande tesauro fonte, ou (ii) fazendo o uso de linguagens de troca, particularmente um léxico intermediário, que é um dispositivo (e não um novo vocabulário construído a partir dos originais, como vemos na figura 1) que vai permitir a tradução de termos de indexação de um esquema de indexação para outro. Para isto, um léxico intermediário precisa apresentar algumas condições básicas em sua construção. Sua estrutura deve ser flexível o suficiente para a inserção de novos conceitos e estes devem ser

definidos sem ambiguidade, além disso o léxico intermediário deve ser o mais abrangente possível para incluir todos os tópicos cobertos pelos esquemas com os quais vai ser usado.

Figura 1 – Léxico intermediário



Fonte: Horsnell (1975)

Também nesta direção, consideramos de grande interesse discutir o conceito de "Sistema de Coordenadas Semânticas", apresentado por Pierre Lévy (LÉVY, 2014, p. 312, 2019, p. 27), cuja formulação e procedimentos metodológicos permitem uma aproximação com o conceito de dispositivos de comutação, e consequentemente com os léxicos intermediários, colocando uma visão atual e centrada em procedimentos computacionais e automáticos para esta implementação.

Pierre Lévy vem se dedicando a explicitar uma construção teórica que ele denomina de IEML – Metalinguagem da Economia da Informação e argumenta que a sua principal hipótese para propor tal metalinguagem é a de que ainda não inventamos sistemas simbólicos que se encaixam no novo meio digital. Ao propor esta construção ele define o IEML como: 1. uma linguagem artificial que se traduz automaticamente em línguas naturais;

2. uma linguagem de metadados para a marcação semântica colaborativa de dados digitais; 3. uma nova camada de endereçamento do meio digital (endereçamento conceitual) resolvendo o problema de interoperabilidade semântica; 4. uma linguagem de programação especializada no design de redes semânticas; 5. um sistema de coordenadas semântica da mente (a esfera semântica), permitindo a modelagem computacional da cognição humana e a auto-observação das inteligências coletivas (LÉVY, 2014).

Num caminho que interessa à nossa discussão sobre os dispositivos intermediários de conversão de linguagens e SOCs, Lévy define que a IEML é uma linguagem formal e uma linguagem natural, que calcula, gera e reconhece automaticamente uma infinidade de conceitos e seus relacionamentos semânticos. Em suma, a IEML não é uma ontologia universal, é uma linguagem com semântica calculável que pode expressar qualquer ontologia (LÉVY, 2019, p. 25). O que o programa de pesquisa da IEML pretende, como afirma Lévy (2009), não é impor uma super ontologia (já que o IEML pode ser usado no contexto de ontologias com hierarquias de conceitos muito diferentes), mas o de fornecer uma linguagem na qual pelo menos as classes mais altas de qualquer ontologia e sistema de classificação, bem como as tags populares produzidas pelo usuário de folksonomias possam ser traduzidas.

Neste sentido, na construção teórica proposta por Lévy, nos interessa sobremaneira os conceitos de linguagem ponte e de um sistema de coordenadas semânticas, apresentadas na IEML. Como linguagem ponte podemos entender uma linguagem artificial ou natural usada como uma linguagem intermediária para tradução entre muitas línguas diferentes – para traduzir entre qualquer par de idiomas A e B, uma traduz A para a linguagem ponte L, depois de L para B. Como sistema de coordenadas semânticas é importante entender que sua função seria de permitir um sistema de endereçamento que possibilite computar as relações e distanciamentos semânticos existentes entre as linguagens (LÉVY, 2014).

4 UM CAMINHO PARA A COMPATIBILIZAÇÃO DE VOCABULÁRIOS MANTENDO OS VOCABULÁRIOS ORIGINAIS

Na construção de uma proposta para interoperar sistemas de organização do conhecimento, sem interferir nos vocabulários originalmente construídos, visando uma recuperação inteligente que possa futuramente atender a um “sistema de coordenadas semânticas”, é necessário se dar um primeiro passo. A partir de todo o estudo realizado e adotando um método de análise comparativa, visando o que melhor poderia se consolidado

em termos da CI e da CC, construímos um caminho de aplicação para estabelecermos uma prova de conceito que pudesse nos auxiliar objetivando a interoperação de ambientes heterogêneos visando uma recuperação inteligente, onde um usuário possa fazer um caminho único de busca de informação.

Para avançar nesta questão podemos dizer ser necessário criar uma identificação única[1] para os conceitos de determinado ambiente e esta identificação deve ser semanticamente representativa da significação (ou da semântica) de cada conceito. Dahlberg (1981) ao propor o seu registro do conceito já tinha esta preocupação, apesar de sua implementação se dar em matrizes de correlação e estar apoiada nos parcos recursos computacionais existentes à época.

Portanto, com o objetivo de chegar a uma recuperação inteligente da informação e resolver os problemas que dificultam a interoperabilidade, o que propomos aqui é a utilização de técnicas da Ciência da Computação para que, com base nos procedimentos propostos por Dahlberg, seja possível construir registros de conceito que funcionem como elementos de comutação entre sistemas de organização do conhecimento, formando os identificadores únicos de cada conceito de um ambiente heterogêneo a ser compatibilizado. A partir dos procedimentos de mapeamento estas ligações seriam estabelecidas entre os vocabulários participantes por meio de medidas de similaridade semântica.

As técnicas oferecidas pela Ciência da Computação que utilizamos em nosso estudo, chamadas de técnicas de correspondência, são aplicadas em nível de elemento e em nível de estrutura e podem ser base para algoritmos computacionais que manipulem as expressões verbais e as cadeias de caracteres dos registros de conceitos gerados extraindo sua expressão semântica, através de medidas de similaridade semântica. As Técnicas em Nível de Elemento utilizam os valores literais dos conceitos, e/ou suas propriedades, para medir a similaridade semântica. Existem cinco técnicas de nível de elemento: baseadas em recursos formais, baseadas em recursos informais, baseadas em strings (cadeias de caracteres), baseadas em linguagem e baseadas em restrições. Além das técnicas no nível de elemento, temos as

[1] Qualquer proposta de tipo de identificador ou localizador de recursos utilizado hoje para ser usado em um ambiente de rede precisa levar em consideração as limitações da Internet. Em que pese a necessidade, neste caso, de avançarmos para identificadores 'transparentes' para recursos na Web, o que temos à disposição são as URLs e URIs, que atuam, com pequenas nuances entre si, como localizadores e são de natureza opaca. Por isso, nossa proposta de implementação é com o uso de URIs.

técnicas de correspondência no nível de estrutura, que usam a estrutura do SOC para calcular a similaridade entre conceitos. Foi possível destacar quatro técnicas no nível de estrutura, ou seja, aquelas baseadas em taxonomia, baseadas em grafos, baseadas em instância e baseadas em modelo (ANGERMANN; RAMZAN, 2017).

A partir de todo o levantamento de literatura realizado, pudemos mostrar que os processos de alinhamento semântico e mapeamento são os procedimentos fundamentais para compatibilizar diferentes SOC (BRUIJN et al, 2006; EHRIG, 2007; EUZENAT; SHVAIKO, 2013). Mas, o que defendemos aqui é que estes procedimentos devem ser utilizados não simplesmente para gerar mapeamentos restritos entre dois vocabulários apenas (que geram uma grande quantidade de mapeamentos um para um no caso de vários vocabulários) e também que não sejam usados para a junção, mesclagem e união de vocabulários. Pudemos ver que os mesmos processos podem ser usados tanto para esta aglomeração de vocabulários quanto para a geração de dispositivos de comutação que permitam chegar a uma proposta de coordenadas semânticas.

Neste sentido, consideramos construir um caminho utilizando as técnicas e métodos apresentados na CI e na CC para colocar em ação um modelo proposto para verificarmos o que mais atenderia ao propósito de criação de um ambiente de coordenadas semânticas, voltado para a recuperação da informação entre bases indexadas por SOC heterogêneos.

5 RESULTADOS

As teorias estudadas permitiram formular uma proposta de procedimentos que, aplicados sobre o recorte escolhido, objetivaram conseguir demonstrar a possibilidade e viabilidade da implementação de procedimentos automáticos capazes de traçar um caminho para a interoperabilidade semântica de sistemas de organização do conhecimento. Este caminho pressupõe que:

(i) os vocabulários originais continuam a ser utilizados para indexar as bases de dados que já indexavam;

(ii) não há necessidade de modificação nos vocabulários originais;

(iii) não há criação de novos vocabulários, mesclados ou aglutinados, a partir dos vocabulários originais;

(iv) utiliza-se as bases teóricas e metodológicas que compreendem procedimentos e técnicas já propostas pela CI e pela CC, mas que sejam combinadas de

forma a gerar um processo automático viável com a utilização das tecnologias da informação atuais;

(v) objetiva-se a criação de um espaço semântico, de inspiração na proposta de coordenadas semânticas de Pierre Lévy e nas propostas de linguagens intermediárias propostas pela CI, que seja capaz de mapear conceitos de forma semântica entre os vocabulários participantes e de municiar um sistema de recuperação da informação que recupere informações distribuídas ao navegar neste espaço semântico e interagir com os utilizadores.

Os termos escolhidos, conforme apontado em nossa metodologia, foram usados para gerar, cada um deles, um registro de conceito com os seguintes campos: (i) nome do conceito; (ii) notação; (iii) termo genérico imediato (ou simplesmente termo genérico); (iv) termo genérico maior; (v) termos específicos; (vi) termos associados.

Desta forma, a construção do espaço semântico de recuperação inteligente de informações se inicia com a construção, de forma automatizada e sem alterar os vocabulários originais, que sempre foi nosso propósito, deste objeto chamado registro de conceito, que permitirá aglutinar as informações referentes a este conceito e, posteriormente, proporcionar a realização de mapeamentos que o liguem a conceitos totalmente ou parcialmente similares em outros vocabulários. E esta é uma questão chave a ser considerada aqui, ou seja, não podemos nos limitar a estabelecer correspondências e mapeamentos onde ocorre apenas a dualidade está ou não-está mapeado, e sim procurar estabelecer caminhos para criar medidas de similaridade, que por sua vez possam ser usadas para prover recursos de recuperação inteligente e interativa entre sistemas de recuperação de informações e usuários, num ambiente de diversos SOC e bases de dados.

Estas possibilidades de correspondência foram geradas quando procuramos responder as seguintes perguntas: (i) como relacionar semanticamente um conceito presente em um vocabulário com conceitos de outros vocabulários? (ii) ao encontrar a presença de conceito com expressão verbal ou semelhante em outro vocabulário, sua compatibilidade semântica pode ser confirmada? (iii) quais outros conceitos em outros vocabulários podem ser mapeados com alguma similaridade semântica ao conceito buscado, mesmo que tenham expressão verbal diferentes?

Para a verificação da validade do experimento, procedemos à aplicação das técnicas e algoritmos computacionais de manipulação de cadeias de caracteres como lematização e

morfologia, tokenização, eliminação de plurais, mostradas por Angermann e Ramzan (2017), além das técnicas que tratam da similaridade de nome e similaridade de descrição, em especial, mas não apenas, a distância de Levenshtein (1966), a distância de Bailey (2004), a distância de Hamming (1950), a similaridade Cosine (TAN; STEINBACH; KUMAR, 2005), a distância de Jaccard (PARK; KIM, 2007), a distância de Lin (KERNIGHAN; LIN, 1970), e as técnicas de navegação na estrutura do vocabulário, tais como as técnicas baseadas em taxonomia, as técnicas baseadas em grafos, as técnicas baseadas em instâncias e baseadas em modelos (CHEATHAM; HITZLER, 2013; ANGERMANN; RAMZAN, 2017). A utilização comum deste conjunto de técnicas exemplificadas aqui permite uma comparação entre conceitos, superando a trivial comparação sintática.

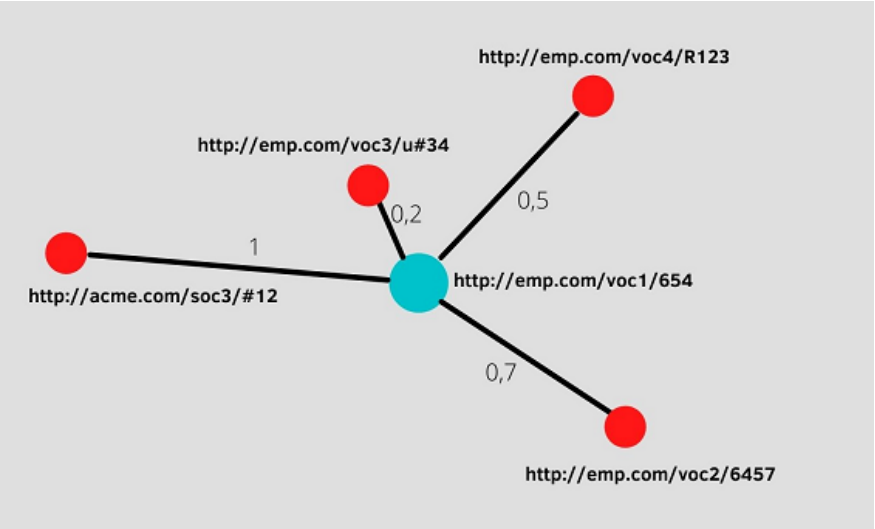
Estas operações se voltam para dois procedimentos básicos, ou seja,

(i) verificar se as formas verbais que se equivalem sintaticamente representam conceitos que são semanticamente iguais ou semelhantes, ou se são polissemias;

(ii) descobrir nos vocabulários participantes possíveis expressões verbais que, apesar de dessemelhantes sintaticamente, apresentam similaridade semântica que as tornem capazes de serem mapeadas dentro do espaço semântico construído.

Estes mapeamentos n-dimensionais compõem o que chamamos de “espaço semântico” (figura 2), que por sua vez compõe a base de um sistema de recuperação inteligente de informações, onde cada mapeamento contém como atributo a distância semântica entre os conceitos mapeados. Nesta figura 2 – Correspondências semânticas a partir de um conceito – podemos ver uma representação esquemática dos conceitos, grafados aqui como círculos e identificados por URIs em cinco diferentes sistemas de organização da informação. A partir do conceito identificado por http://emp.com/voc1/654 (círculo azul) é possível verificar que este conceito tem, por exemplo, uma medida de similaridade semântica 1 com o conceito http://acme.com/soc3/#12, além de diferentes medidas de similaridade calculadas com outros conceitos (círculos vermelhos). Através da navegação num espaço semântico deste tipo podemos avançar para a criação de sistemas de recuperação da informação capazes de recuperar informações em diferentes SOC de forma inteligente, ou seja, a partir das correlações semânticas entre os conceitos e não de sua expressão verbal. Num ambiente deste tipo a expressão verbal do conceito é apenas um dos pontos de comparação e análise para estabelecer as distâncias semânticas.

Figura 2 – Correspondências semânticas a partir de um conceito



Fonte: Elaborado pelo autor

No experimento realizado em nosso campo empírico foi possível verificar que em 94% dos termos estudados, incluindo aqueles cuja expressão verbal ocorreu em dois ou apenas um dos três vocabulários estudados, pudemos estabelecer mapeamentos que já podemos chamar de semânticos, pois traduzem uma identidade semântica de cada conceito a partir de sua identificação como registro de conceito.

O que apresentamos aqui, portanto, como caminho a ser seguido não é o estabelecimento de uma matriz de mapeamento booleano, onde a ligação entre conceitos existe ou não existe. O caminho apresentado é o estabelecimento de uma base de dados de registros de conceitos, representando um grupo de SOCs participantes que, ao ser gerada a partir dos caminhos propostos aqui vai ser utilizada por um sistema de recuperação inteligente da informação, oferecendo uma visão semântica do conjunto de sistemas de organização do conhecimento compatibilizados e servindo de base para buscas interativas inteligentes entre os diversos SOCs por parte dos usuários do sistema.

6 CONSIDERAÇÕES FINAIS

Nossa proposta foi apresentar aqui um caminho possível e viável de ser implementado utilizando os recursos e a inteligência que já estão presentes nos sistemas de organização do conhecimento tradicionais que, aliados às técnicas e procedimentos computacionais que já temos disponíveis e as bases teóricas da Ciência da Informação, possam apresentar uma proposta para a solução da heterogeneidade. Este caminho pode, por um lado, resolver problemas e compatibilidade com vocabulários já existentes, mas pode também, ser usado para orientar e propor possibilidades de construção mais adequadas para SOC ainda a serem desenvolvidos, que utilizem novas tecnologias.

Ainda com relação aos resultados obtidos de nossos procedimentos empíricos, foi possível observar que nossas premissas da possibilidade de avançar na criação de um espaço semântico se justificaram. Os exercícios de compatibilização permitiram a realização de mapeamentos semânticos em mais de 90% dos termos utilizados em nosso experimento, conforme explicitado acima.

Por fim, com base nos estudos e experimentos realizados fomos capazes de cumprir nosso objetivo de apontar um caminho que possa ser capaz de ser aplicado a vocabulários de ambientes heterogêneos e gerar um espaço semântico. Desta forma, propomos que, a partir daí, este dispositivo possa ser utilizado por um SRI que permita a recuperação inteligente da informação nestes ambientes, de forma flexível, permitindo sua expansão, e interativo, que permita a participação do usuário de forma ativa nas suas buscas.

Diversos estudos futuros podem ser realizados com a continuação e ampliação do presente estudo, entre eles o estabelecimento de critérios para a determinação das medidas matemáticas de similaridade semântica a serem anotadas entre os conceitos, a elaboração de um protótipo de agente de software que implemente de forma automática os experimentos aqui realizados e a codificação de um SRI que faça uso do espaço semântico aqui proposto.

REFERÊNCIAS

AGROVOC. **Multilingual Thesaurus**. 2021. Disponível em: http://www.fao.org/agrovoc/search. Acesso em 17 mai. 2021.

ANGERMANN, H.; RAMZAN, N. **Taxonomy Matching Using Background Knowledge**: Linked Data, Semantic Web and Heterogeneous Repositories. Springer International Publishing. 2017.

BAILEY, D. An efficient euclidean distance transform. In: 11th International Semantic Web Conference (Springer). **Proceedings** […]. Berlin, Germany. p. 394–408. 2004.

BRUIJN, J. et al. **Ontology mediation, merging and aligning**. May 20. 2006.

CABI. **CAB International**. 2019. Disponível em: https://www.cabi.org/cabthesaurus/mtwdk.exe?yi=home. Acesso em 17 mai. 2021.

DAHLBERG, I., Towards establishment of compatibility between indexing languages. **International Classification**, v. 8, n. 2, p. 88-91, 1981.

EHRIG, M. **Ontology Alignment**—Bridging the Semantic Gap. Semantic Web and Beyond, Springer, New York. 2007.

EUZENAT, J.; SHVAIKO, P. **Ontology Matching**. 2. ed. Springer, Heidelberg. 2013.

HAMMING, R. W. Error detecting and error correcting codes. **Bell System Technical Journal**, 29 (2): 147–160. 1950. Disponível em: https://signallake.com/innovation/hamming.pdf. Acesso em 17 mai. 2021.

HORSNELL, V. The Intermediate Lexicon: an aid to international co-operation. **Aslib Proceedings** 27 (2), February 1975, pp. 57-66. 1975.

KERNIGHAN, B.; LIN, S. An efficient heuristic procedure for partitioning graphs. **Bell Syst. Tech. J.** (Blackwell Publishing) 49, 291. 1970.

LEVENSHTEIN, W. Binary codes capable of correcting deletions, insertions and reversals. **Sov. Phys. Dokl**. (Springer) 10, 707–710. 1966.

LÉVY, P. **A Esfera Semântica**. Tomo 1: Computação, cognição, economia da informação. Editora Annablume. 2014.

LÉVY, P**. From Social Computing to Reflexive Collective intelligence**: The IEML Research Program. CRC, FRSC, University of Ottawa. 2009.

LÉVY, P. **IEML - A metalinguagem da Economia da Informação** - Livro Branco. Pré-print, não publicado. DOI: 10.13140/RG.2.2.11232.33281. 2019.

PARK, S.; KIM, W. Ontology mapping between heterogeneous product taxonomies in an electronic commerce environment. **Int. J. Electron. Commer**. (M. E. Sharpe) 12, 69–87. 2007.

TAN, P. N.; STEINBACH, M.; KUMAR, V. **Introduction to data mining**. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2005. ISBN 0321321367.

THESAGRO. **Sistema especializado em literatura agrícola**. 2020. Disponível em: http://sistemas.agricultura.gov.br/tematres/vocab/index.php. Acesso em: 17 mai. 2021.