MULTÍPLOS OLHARES EM CIÊNCIA DA INFORMAÇÃO
Vol. 9, Nº. 2 Ano 2019

COMPARAÇÃO DE POSSIBILIDADES DE RECUPERAÇÃO DA INFORMAÇÃO EM TERMINOLOGIAS BIOMÉDICAS

Comparison of Information Retrieval Possibilities in Biomedical Terminologies

Eduardo Ribeiro Felipe
Programa de Pós-graduação em Gestão e Organização do Conhecimento, Universidade Federal de Minas Gerais - UFMG
erfelipe@ufmg.br

Maurício Barcellos Almeida
Programa de Pós-graduação em Gestão e Organização do Conhecimento, Universidade Federal de Minas Gerais - UFMG
mba@eci.ufmg.br

RESUMO

A representação da informação passou por muitos estágios até chegar onde a humanidade situa-se agora, em um ambiente em que prevalece a representação digital. Período marcado pela grande produção de informação, e também da necessidade sua recuperação com precisão. Neste contexto, a publicação de artigos científicos vem aumentando significativamente no ambiente acadêmico. Este tem sido o meio mais utilizado para comunicação científica e recebe cada vez mais destaque nos modelos de reconhecimento tanto em trabalhos na academia, bem como em iniciativas de desenvolvimento prático no mercado de trabalho. Ao considerar esta perspectiva, instrumentos de organização e representação do conhecimento tornam-se importantes também na Recuperação da Informação a partir da possibilidade na expansão de *querys*, que pode ser um método para a pesquisa por artigos científicos pertencentes à mesma temática que o instrumento representa.

Palavras-Chave: Recuperação da Informação; Ontologia; Tesauro; Expansão de consulta; Correspondência textual.

ABSTRACT

Information representation has gone through many stages until it reaches where humanity now stands, in an environment where digital representation prevails. Period marked by the great production of information, and also the need for its recovery with precision. In this context, the publication of scientific articles has been increasing significantly in the academic environment. This has been the most widely used medium for scientific communication and is increasingly highlighted in recognition models in both academic work and practical development initiatives in the labor market. Considering this perspective, knowledge organization and representation tools also become important in information retrieval based on the possibility of queries expansion, which can be a method for researching scientific articles on the same theme that the instrument represents.

Keywords: Information Retrieval; Ontology; Thesaurus; Query expansion; String match.

Universidade Federal de Minas Gerais

Correspondência/Contato
Av. Antônio Carlos, 6627
Pampulha: 31270-901
BELO HORIZONTE - MG

Escola de Ciência da Informação da UFMG

1. INTRODUÇÃO

A transmissão de conhecimento é uma característica fundamental da espécie humana. A capacidade dos seres humanos em representar a informação, produzir sua codificação, decodificação, armazenamento e recuperação são fatores que permitiram o avanço da ciência e da sociedade como vemos hoje, caracterizada pela importância da organização da informação e do conhecimento.

No ambiente acadêmico, a importância da informação surge de um modo ampliado. Pode-se considerar a informação gerada pela metodologia aplicada como o produto e o resultado da pesquisa. O modo como este resultado é divulgado na comunidade científica é representado em publicações científicas. Esse formato é utilizado há muito tempo e permanece inalterado (STOCKER *et al*., 2018), além de ser considerado o mais importante veículo de comunicação científica.

Embora o surgimento do "digital" na segunda metade da década do século XX tenha trazido enormes avanços na comunicação humana e na produção de informação, vários processos ainda carecem de tratamento e melhores soluções. Há diversos desafios relacionados ao processamento da informação, desde sua representação, armazenamento e recuperação, sendo este último, o foco deste estudo. Considerando o atual período, onde a produção de informação excede em muito a capacidade humana de organizá-la, para fins de recuperação plena e satisfatória, os algoritmos computacionais ainda precisam evoluir para processar dados em múltiplos formatos como textos, figuras, símbolos, tabelas e gráficos.

A publicação de artigos científicos tem crescido exponencialmente pois muitas instituições traçam como parâmetro de êxito o número de publicações de artigos, bem como o crescente número de bases de dados que competem na busca por melhores rankings de reconhecimento e reputação (BJÖRK; ROOS; LAURI, 2008). No Brasil esse reconhecimento é responsabilidade do Ministério da Educação e suas iniciativas como a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES[1]), uma fundação governamental, a qual avalia os periódicos em uma escala classificatória por meio do modelo QUALIS[2].

Os artigos, portanto, são o principal instrumento de transmissão do conhecimento acadêmico e recebem, a cada dia, maior ênfase pelas instituições. Estas últimas,

[1] https://www.capes.gov.br/pt/historia-e-missao
[2] https://sucupira.capes.gov.br/sucupira/public/index.xhtml

por sua vez, recebem status e recursos pelo número de publicações de seu corpo docente e discente.

Ao mesmo tempo que se percebe a valorização do artigo enquanto instrumento de preservação e transmissão do conhecimento científico, a área da Recuperação da Informação (RI) recebe de igual forma, uma importância crescente. Tanto na Ciência da Computação (CC) quanto na Ciência da Informação (CI) há um crescente interesse pelos temas relacionados a Banco de Dados, Sistemas de Informação, Bases de Conhecimento bem como Sistemas Integrados (AAMODT; NYGÅRD, 1995). Considerando ainda essas duas áreas de conhecimento, tanto a CC como CI tratam sobre o tema Ontologia, instrumento para representação do conhecimento e que é relevante para os objetivos deste trabalho. Almeida (2013) explica que ontologia na CC é utilizada para referir-se a um vocabulário em linguagem de representação do conhecimento e também como uma estrutura teórica que explica fenômenos por meio de fatos e regras. Na CI, ontologias também podem ser usadas para a construção de estruturas de categorias na representação de conteúdo documental. Farinelli e Almeida (2019) conceituam que uma ontologia descreve o significado dos símbolos adotados no sistema de informação e representa uma visão específica do mundo e Mendonça *et al.* (2015) afirmam que no campo da Web Semântica e da engenharia ontológica, a conceitualização é considerada uma pedra angular. Destaca-se ainda que a CI possui outros instrumentos voltados ao processo informacional como as listas de cabeçalho de assunto, taxonomias e tesauros, este último melhor detalhado adiante neste trabalho.

2. CONTEXTO

O interesse pela temática desta pesquisa deu-se pela constatação de que, instrumentos de organização e representação do conhecimento possuem grande amplitude terminológica e conceitual, ou seja, são capazes de fornecer conexões conceituais de grande expressividade através de seus termos e relações.

Ao considerar esta perspectiva, estes instrumentos tornam-se importantes também na Recuperação da Informação a partir da possibilidade na expansão de *querys*[3], que pode ser um método para a pesquisa por artigos científicos pertencentes à mesma temática que o instrumento representa.

[3] Consulta em uma estrutura de dados para obtenção de resultados

Dessa forma, o presente trabalho enquanto pesquisa no campo da informação, tem como objetivo geral investigar a revocação de artigos científicos no processo de recuperação da informação utilizando dois instrumentos de organização e representação do conhecimento da área médica, as Terminologias Clínicas: *Systematized Nomenclature of Medicine - Clinical Terms* (SNOMED CT) e o *Medical Subject Headings* (MeSH). Os objetivos específicos são: a) construção de um banco de dados para persistência dos artigos científicos, b) construção de algoritmos para realização de *querys* expandidas com relações hierárquicas e axiomáticas na base de dados, c) submissão das *querys* ao banco de dados, d) comparação de resultados estatísticos.

3. FUNDAMENTAÇÃO

A Recuperação da Informação está intimamente ligada aos processos de representação, armazenamento, organização e acesso a recursos de informação. O conjunto de itens que pode ser processado na RI é vasto e pode ser exemplificado por cartas, documentos, jornais, artigos, livros, prontuários médicos, entre outros (SALTON; MCGILL, 1983). Segundo Vickery (1970), "o problema da recuperação da informação não é algo novo. Ferramentas para identificar documentos [...] existem há séculos, mas o termo recuperação da informação raramente foi encontrado antes de 1955." De fato, o termo Recuperação da Informação teve seu primeiro uso por Calvin Mooers (1950), na observação do processo de recuperação digital da informação. Seu trabalho, já naquela época, apontava para os desafios para recuperar informação "não numérica", onde a construção não estruturada da linguagem natural, a semântica e as grandes coleções, resultado do crescimento da pesquisa científica, eram desafios visíveis. Neste ponto, cabe citar literatura sólida da área disponível a partir de Salton e Mcgill (1983), Chowdhury (2004), Baeza Yates e Ribeiro Neto (2011), dentre outros.

O trabalho de RI exige, em grande parte de suas aplicações, a tradução entre a linguagem do usuário e a usada pelos autores na identificação do documento. Esta tradução é feita por vocabulários controlados baseados em uma área de domínio do conhecimento, usados como uma ponte entre a terminologia na construção da query com os termos usados na representação documental (HOPPE; HUMM; REIBOLD, 2018). O desafio no equilíbrio entre revocação e precisão é outro desafio sempre presente nos projetos de RI. Frické (2012) afirma que revocação é a presença de sinal, precisão é a ausência de ruído. E o que se procura é um sinal forte alinhado a um baixo ruído. Essa

busca pelo equilíbrio ideal em RI torna-se mais complexa com a consideração da semântica. Por outro lado, o uso da semântica na recuperação permite que um maior número de conceitos relacionados seja recuperado (MUKHERJEA, 2005), ampliando os horizontes para o usuário na identificação de documentos que não estariam presentes em uma recuperação tradicional.

A ciência da informação há décadas vem contribuindo no processo de recuperação da informação, com ênfase em mecanismos que possam viabilizar tanto a organização informacional, quanto a sua recuperação. É parte de sua vocação enquanto área do conhecimento. Borko (1968) define que "a ciência da informação é a disciplina que investiga as propriedades e o comportamento da informação, as forças que governam o fluxo de informações e os meios de processar informações para otimizar a acessibilidade e a usabilidade", e Saracevic (1979) entende que "o problema básico abordado na ciência da informação é com a eficácia na comunicação do conhecimento público".

Dos diversos instrumentos utilizados pela CI, destacam-se neste trabalho os tesauros, vocabulários controlados que tratam de um sistema linguístico no qual os componentes principais são os termos (CURRÁS, 1995). Entende-se por "termos" a definição de Felbert (*apud* CURRÁS, 1995, p. 77): "unidades linguísticas de um vocabulário especializado". Este instrumento possui como características uma organização em estrutura hierárquica, dispondo os termos logicamente, permitindo estabelecer conexões semânticas entre os termos na eleição de termos "principais" e termos relacionados.

São tipos de relações terminológicas em tesauros:

- BT: *broader term*
- NT: *narrower term*
- RT: *related term*
- SN: *synonym*
- USE: *use instead*
- UF: *used to mean*

Todavia, a construção de um tesauro exige o conhecimento e experiência do profissional da informação. A escolha dos termos e suas relações é o que define a qua-

lidade, a abrangência de utilização do instrumento em sua adoção prática, tanto na fase de classificação / indexação, quanto na fase de recuperação informacional.

Estes instrumentos terminológicos são usados na tradução entre a linguagem (terminologia) usada pelo usuário em contraste com a terminologia da base de conhecimento. São, portanto, construídos com o foco na transição da linguagem natural, com sua característica de proximidade ao discurso direto, para a linguagem estruturada. No contexto deste trabalho, a terminologia da base de conhecimento é um vocábulo médico.

O processo histórico dos tesauros possui uma boa referência no trabalho de Mendes, Reis e Maculan (2015) e Currás (1995). Vários autores convergem no entendimento que este instrumento é uma espécie de evolução se comparado aos cabeçalhos de assunto, estruturas terminológicas que não possuem hierarquia e conexões semânticas. Sua importância na RI possui relevância contínua, mesmo em cenários contemporâneos, onde são desenvolvidos métodos estatísticos de recuperação de texto completo ou aplicações que usam lógica formal associadas à pesquisa na web semântica (TUDHOPE; BINDING, 2016).

Vieira, Santos e Lapa (2010), por exemplo, reforça em seu trabalho o papel do tesauro no processo de indexação. Nesta abordagem, há um claro direcionamento em apontar o tesauro como fonte de termos descritores na representação documental. Deste modo, mesmo que a RI seja composta por um conjunto de processos que pode ser abordado em grandes grupos como: indexação, gravação e recuperação; alguns autores enfatizam as linguagens documentárias a exemplo dos tesauros, como principais fornecedores de descritores na indexação. Isso parece configurar uma visão reducionista do potencial que o instrumento pode exercer na RI.

4. METODOLOGIA

A pesquisa, ainda em andamento, pode ser classificada da seguinte forma segundo Gil (2010): segundo a área de conhecimento, há uma intrínseca interseção e interdisciplinaridade entre as áreas de Ciências Sociais Aplicadas e de Ciências Exatas e da Terra. Quanto à sua finalidade trata-se de uma pesquisa aplicada. Quanto aos seus objetivos, é considerada exploratória, pois envolve o levantamento bibliográfico das temáticas pertinentes e o desenvolvimento de algoritmos para a experiência aplicada. Quanto à natureza dos dados, é uma pesquisa quantitativa. Quanto ao ambiente em que estes

dados são coletados, trata-se de uma pesquisa de laboratório. Como modalidade, a pesquisa é um estudo de caso, por realizar uma aplicação de situação de experimento prático na condução da análise de dados.

O estudo de caso pretende ser viabilizado em etapas que demandam os seguintes aspectos técnicos e tecnológicos:

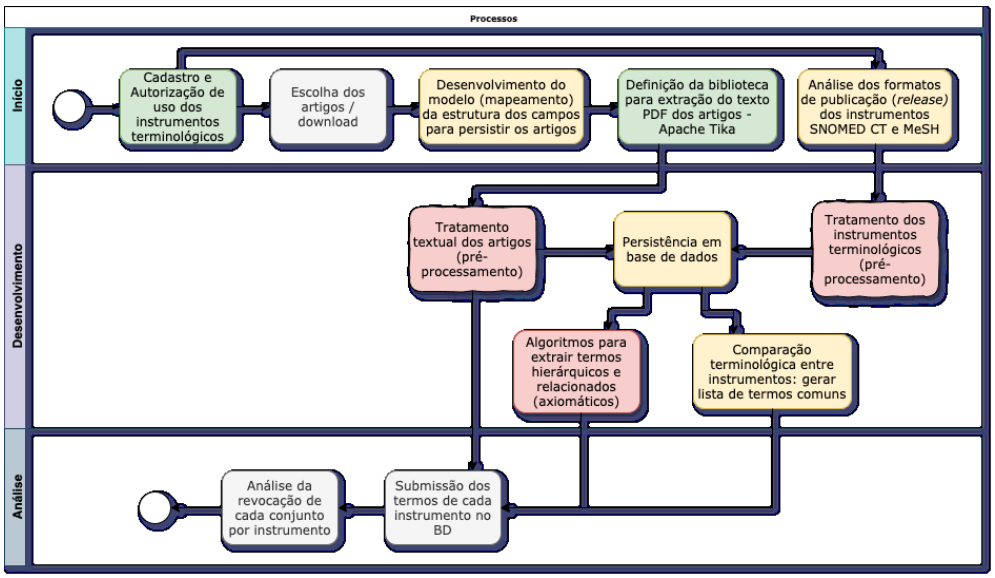

Figura 1. Processos do estudo de caso
Fonte: Elaborado pelos autores.

4.1. Obtenção dos instrumentos terminológicos

Após a análise para definição dos instrumentos terminológicos mais adequados para uso no projeto de pesquisa, dois foram escolhidos: *Systematized Nomenclature of Medicine - Clinical Terms* (SNOMED CT) e *Medical Subject Headings* (MeSH), foram os selecionados a partir dos seguintes critérios:

i. Credibilidade;

ii. Abrangência semântica em diversas áreas do conhecimento médico;

iii. Disponibilidade;

iv. Política de atualização.

O procedimento para acesso e uso dos instrumentos foi realizado em meio digital, iniciado no preenchimento de cadastros no site de cada instituição. Para o SNOMED CT há uma licença denominada *Member Licensing and Distribution Service* (MLDS), que pode ser requisitada por um formulário eletrônico a ser analisado pela instituição publicadora. Após a aprovação, uma área de gerenciamento acessada por

browser é disponibilizada, bem como *login* e senha para acesso. Há no mínimo duas atualizações durante o ano, que podem ser acessadas diretamente no site, em formato compactado para download, bem como diversas documentações. No *Medical Subject Headings* (MeSH), o processo de solicitação de acesso é similar e também conta com uma análise para autorização após o preenchimento do formulário eletrônico. A instituição *U.S. National Library of Medicine* é a responsável e sua política de disponibilização inclui atualizações semanais e diversos formatos de distribuição.

4.2. Escolha e obtenção dos artigos científicos para análise

Inicialmente foram definidas duas áreas temáticas, à saber, Obstetrícia ou Geriatria, as quais são ligadas ao grupo de pesquisa: Representação do Conhecimento, Ontologias e Linguagem (Recol[4]), do qual os autores são participantes. A escolha dos artigos do acervo da *BioMed Central Ltd* (BMC) foi decidida pelos seguintes fatores:

i. Disponibilidade temática;

ii. Publicação dos artigos em formato completo (*full papers*);

iii. Acervo de livre acesso (*open access*);

iv. Instituição com credibilidade no meio acadêmico;

v. Quantidade de artigos disponíveis.

A BMC é filiada da *Springer Nature,* outra instituição de grande reconhecimento científico. A organização de seus artigos está agrupada por classificação de grandes temas, e dentre os temas selecionados para este trabalho, a temática "Geriatria" foi encontrada disponível. Dessa forma, foi realizado o processo de download dos artigos, totalizando neste momento do trabalho 2.364 arquivos em formato *Portable Document Format* (PDF). Os processos para extração do texto, identificação de metadados e pré-tratamento serão descritos no tópico de tratamento da informação.

4.3. Definição e criação da base de dados de artigos

Para permitir o manuseio dos artigos em um formato passível de pesquisa, concluiu-se por meio de análise que uma estrutura de banco de dados seria adequada para estrutu-

[4] http://recol.eci.ufmg.br/

rar a informação extraída dos arquivos PDF e permitir a submissão de *querys* para obtenção dos dados estatísticos.

O banco de dados (BD) escolhido foi o produto *ElasticSearch* (ES). Este produto é um *software* livre, de grande impacto na comunidade por oferecer flexibilidade na constituição de estruturas de dados, bem como sua alta performance. Trata-se de um banco de dados NoSQL[5], orientado a documentos, voltado para o ambiente web com interface de comunicação *Representational State Transfer* (REST[6]), o que permite o acesso aos dados nele persistidos por várias linguagens de programação. A estrutura documental (que remete às tabelas no modelo relacional) foi definida utilizando campos do modelo *Dublin Core*[7]. Neste modelo serão persistidos os dados extraídos dos artigos originalmente em PDF.

4.4. Extração dos dados em formato PDF para texto sem formatação

Esta etapa possui grande importância por lidar diretamente com os dados sob pesquisa. Uma abordagem inadequada no processamento textual reflete instantaneamente no processo de recuperação, foco deste estudo. Os dados em formato PDF devem ser extraídos e estruturados para permitir a pesquisa e recuperação. Esta extração pode ser feita por meio de bibliotecas de *software* que conseguem acessar o arquivo em formato PDF para recuperar o texto e seus metadados (título, autor, etc...) quando disponíveis. Após vários testes com diferentes bibliotecas, visando a maior fidelidade sintática, a opção escolhida foi a *Apache Tika*[8]. Esta biblioteca foi construída como um subprojeto da Apache Lucene, um famoso projeto desenvolvido em linguagem Java para pesquisa textual em seus processos de tratamento, indexação, persistência e recuperação.

4.5. Tratamento dos instrumentos terminológicos

Os instrumentos selecionados para a implementação do projeto, tanto o SNOMED CT quanto o MeSH possuem particularidades que precisaram ser tratadas antes de sua utilização. Para facilitar o entendimento e elencar suas particularidades, os instrumentos serão explicados separadamente:

[5] Tecnologia que não utiliza *Estructure Query Language* (SQL) e permite esquemas flexíveis para a definição de dados.
[6] Uma arquitetura que oferece requisições interoperáveis para acesso e manipulação de representação de dados via web services.
[7] Padrão de metadados para descrição de objetos digitais.
[8] https://tika.apache.org/

4.5.1. SNOMED CT

Após a aprovação e emissão de um número de filiação, a autorização para uso do SNOMED CT permite o acesso ao instrumento utilizando um painel de controle[9] no ambiente web. A distribuição (*releases*) de cada publicação é listada por data, acompanhada de uma descrição. A versão completa é denominada *SNOMED CT International Edition* e tem a periodicidade semestral. O *download* é feito em formato compactado (ZIP) e após a conclusão do processo e sua descompactação, o formato do instrumento é distribuído em *Release Format 2* (RF2), um conjunto de arquivos texto (TXT) cujos dados, separados por tabulação (TSV[10]), expressam um modelo relacional das informações.

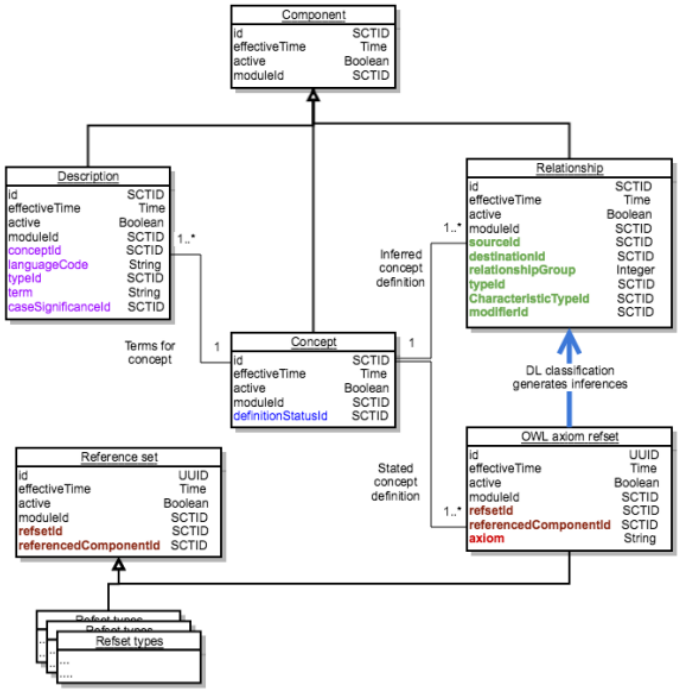

Figura 2. Documentação oficial SNOMED CT – *Release Files Formats* (p. 46)

Nesta pesquisa optou-se por importar as informações para um banco de dados relacional SQLite, de modo a permitir uma manipulação mais prática e eficiente, usando linguagem de consulta SQL[11]. No momento da importação foram processados algoritmos para preparação dos termos (pré-processamento), retirando caracteres especiais, símbolos, operadores lógicos junto à terminologia e inadequações como multiplicidade de termos separados por vírgula, em um mesmo campo.

[9] https://mlds.ihtsdotools.org/#/dashboard
[10] Tab Separated Values
[11] Structure Query Language – Linguagem para consulta e manipulação de informações em banco de dados relacionais

4.5.2. MeSH

O instrumento MeSH também possui um procedimento para cadastro e autorização, mas sua publicação dos dados é muito diferente. Há diversos formatos disponíveis para *download* via *File Transfer Protocol* (FTP) no endereço nlmpubs.nlm.nih.gov, as opções estão em XML, ASCII, MARC 21, RDF. E ainda na modalidade *download,* o pacote *Unified Medical Language System* (UMLS) possui uma versão do MeSH. Mas existem ainda outras opções de acesso ao instrumento como: consulta *on-line* por interface em browser, SPARQL *endpoint* ou RDF *lookup service*. Suas atualizações são diversificadas por formato e alguns possuem alteração diária.

Novamente neste trabalho foi realizado um processo de importação do instrumento para um banco de dados relacional SQLite. Mas não houve aproveitamento do algoritmo do outro instrumento, visto a grande diferença estrutural das terminologias e seus formatos de distribuição. Para manipular a terminologia MeSH, o algoritmo percorre o arquivo em formato XML para extração dos dados e grava as informações pertinentes no banco de dados. Este modelo permite uma manipulação ágil, permitindo selecionar termos e sua estrutura hierárquica usando linguagem de consulta estruturada SQL.

4.6. Identificação dos termos comuns em ambos instrumentos de organização e representação da informação

Embora os instrumentos pertençam à mesma temática e área do conhecimento, a quantidade de termos entre os mesmos é muito díspar. Neste momento da pesquisa o instrumento SNOMED CT conta com 607.976 conceitos e 2.615.178 descrições (termos), enquanto o MeSH é constituído de 207.538 termos. Uma comparação considerando toda a extensão de ambos instrumentos não seria coerente e colocaria uma estatística de análise das relações conceituais muito desigual em sua concepção. Portanto, definiu-se um recorte terminológico para a análise dos termos, usando como critério a igualdade sintática do termo em ambos os instrumentos. Um algoritmo percorre cada um dos termos a partir do SNOMED buscando sua correspondência sintática no MeSH. Uma lista será usada para gravar os termos comuns que estão em ambos os instrumentos. Ela será usada na etapa de submissão de *querys* expandidas, onde uma pesquisa será composta por seus termos hierarquicamente descendentes e quando disponível, suas conexões axiomáticas.

4.7. Extração dos termos hierárquicos e axiomáticos

A estrutura hierárquica é a característica marcante em ambos instrumentos terminológicos escolhidos neste trabalho. A hierarquia permite estabelecer relações entre os termos, indicando por exemplo, agrupamento semântico sob determinado assunto. Em instrumentos como tesauros, esta relação pode ser mais precisa, definindo equivalência ou associação (CURRÁS, 1995).

Este trabalho considera como primeira estratégia de expansão da pesquisa na recuperação de artigos científicos da área médica, a inclusão de termos hierarquicamente descendentes do termo escolhido, caracterizando assim uma especificidade maior na recuperação dos documentos. Dois algoritmos foram desenvolvidos para conter as especificidades de cada instrumento, permitindo a recuperação de seus subtermos de acordo com suas particularidades nas estruturas de dados. Portanto, dado um determinado termo, o algoritmo retorna os respectivos termos descendentes para o enriquecimento da revocação daquela busca.

No caso do SNOMED CT, seus mantenedores a identificam como uma ontologia. De fato, a terminologia possui definições de axiomas em sintaxe[12] normalizada pela W3C para muitos dos termos em sua estrutura. Entretanto, não há consenso no ambiente acadêmico sobre o uso pela terminologia de práticas de modelagem fundamentadas em princípios filosóficos.

Freitas, Schulz e Moraes (2009) explicam que axiomas são sentenças sempre verdadeiras no âmbito de uma área. Almeida (2014) afirma que são estruturas que compõem o formalismo no escopo das ontologias para restringir modelos, de forma a igualar, tanto quanto possível, os modelos que contém o significado pretendido. Estas expressões permitem estabelecer vínculos semânticos entre termos, usando *Object Property*[13], em tradução livre: *propriedade do objeto*. Este recurso permite formar expressões de características do objeto (termo) representando relacionamentos definidos em pares de termos.

4.8. Análise estatística da revocação

Após os processos anteriores em organização e tratamento da informação, é possível realizar a pesquisa dos termos comuns entre os instrumentos no banco de dados de ar-

[12] https://www.w3.org/TR/owl2-syntax/#axioms
[13] https://www.w3.org/TR/owl2-syntax/#Object_Property_Expressions

tigos. Forma-se, portanto, um conjunto terminológico a partir de cada termo da lista, o qual é constituído do termo e seus termos descendentes. No instrumento SNOMED CT, para cada termo do conjunto, pode haver o acréscimo dos termos relacionados pela declaração de seu respectivo axioma.

Este conjunto de dados estatísticos, gerado pela revocação de cada grupo terminológico deve ser analisado em suas dimensões quantitativas, levando em consideração as estruturas hierárquicas de ambos os instrumentos, bem como as possibilidades de expansão pelas declarações axiomáticas quando estiverem declaradas.

5. DISCUSSÃO

No decorrer dos processos citados, as fases que exigiram maior esforço foram as de pré-processamento, tanto dos artigos quanto dos instrumentos terminológicos. Nos artigos o processo de extração levou em conta várias bibliotecas de extração do PDF, as quais acessam o conteúdo proprietário do formato original e o transforma em texto puro. Mesmo a biblioteca selecionada, com performance e qualidade superior às outras, não foi possível realizar o processo livre de falhas.

Pode-se citar como principais problemas a ausência do caractere espaço entre palavras de forma aleatória e a composição do hífen, seguido do símbolo de quebra de linha. Esta última questão foi solucionada por meio de algoritmo que percorre o texto e identifica o símbolo hífen seguido de quebra de linha. O tratamento textual dos artigos ainda incluiu a remoção de caracteres especiais como pontuação, símbolos de parênteses, chaves, colchetes, operadores matemáticos, abreviações a exemplo de "fig.", urls, e o excesso de espaços entre palavras. Estas características, se presentes na notação, são incompatíveis com a sintaxe dos instrumentos terminológicos. Estes por sua vez também receberam tratamento textual de forma a retirar símbolos e operadores lógicos que figuravam na descrição do termo. Uma característica que pode ser questionada é a diversidade da equipe que compõe o grupo de desenvolvimento terminológico. Outro problema potencial nos instrumentos foi a identificação de vários termos separados por vírgula em um mesmo campo de notação. Para contornar essa situação foi considerado o primeiro termo antes do separador, desconsiderando as demais referências.

Espera-se com este trabalho identificar estratégias melhores de recuperação da informação, onde instrumentos terminológicos possam atuar de maneira efetiva, expandindo conceitos e possibilitando ao usuário uma experiência mais rica e precisa.

AGRADECIMENTOS

À Escola de Ciência da Informação da Universidade Federal de Minas Gerais, centro de excelência e promoção do conhecimento, na pessoa do meu orientador, professor Doutor Maurício Barcellos Almeida, cuja trajetória de dedicação e respeito à pesquisa é um grande exemplo. Às pareceristas Dra. Lívia Marangon Duffles Teixeira e Dra. Fernanda Farinelli pelas contribuições e sugestões para a versão final do artigo.

REFERÊNCIAS

AAMODT, A.; NYGÅRD, M. Different roles and mutual dependencies of data, information, and knowledge — An AI perspective on their integration. **Data & Knowledge Engineering**, v. 16, n. 3, p. 191–222, set. 1995.

ALMEIDA, M. B. Revisiting ontologies: A necessary clarification. **Journal of the American Society for Information Science and Technology**, v. 64, n. 8, p. 1682–1693, ago. 2013.

ALMEIDA, M. B. Uma abordagem integrada sobre ontologias: Ciência da Informação, Ciência da Computação e Filosofia. **Perspectivas em Ciência da Informação**, v. 19, n. 3, p. 242–258, set. 2014.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern information retrieval**: the concepts and technology behind search. 2nd ed. New York: Addison Wesley, 2011.

BJÖRK, B.-C.; ROOS, A.; LAURI, M. Global annual volume of peer reviewed scholarly articles and the share available via different Open Access options.In: ELPUB 2008 Conference on Electronic Publishing. **Proceedings…** p. 11, 2008.

BORKO, H. Information science: what is it? **American Documentation**, Washington, v. 19, n. 1, p. 3-5, jan. 1968.

CHOWDHURY, G. G. **Introduction to modern information retrieval**. 2nd ed. London: Facet, 2004.

CURRÁS, E. **Tesauros, linguagens terminológicas**. [s.l.] Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), 1995.

FARINELLI, F.; ALMEIDA, M. B. **Ontologias biomédicas**: teoria e prática. In: Simpósio Brasileiro de Computação Aplicada à Saúde – SBCAS, 18 º, 2019.p. 49.cap.3.

FREITAS, F.; SCHULZ, S.; MORAES, E. Pesquisa de terminologias e ontologias atuais em biologia e medicina. **RECIIS**, v. 3, n. 1, p. 239/248, 11 mar. 2009.

FRICKÉ, M. **Logic and the Organization of Information**. New York, NY: Springer New York, 2012.

GIL, A. C. **Como elaborar projetos de pesquisa**. São Paulo: Atlas, 2010.

HOPPE, T.; HUMM, B.; REIBOLD, A. **Semantic applications**. New York, NY: Springer Berlin, Heidelberg, 2018.

MENDES, P. R.; REIS, R. M. ; MACULAN, B. C. M. S. Tesauros no acesso à informação: uma retrospecção. **Revista ACB: Biblioteconomia em Santa Catarina**, v. 20, n. 1, p. 18, 2015.

MENDONÇA, F. M. *et al*. From a Consensual Conceptual Level to a Formal Ontological Level. In: The Ninth International Conference on Advances in Semantic Processing. **Proceedings**… 2015, p. 6, .

MOOERS, C. N. **The theory of digital handling of non-numerical information and its implications to machine economics**. Boston: Zator Co., 1950.

MUKHERJEA, S. Information retrieval and knowledge discovery utilising a biomedical Semantic Web. **Briefings in Bioinformatics**, v. 6, n. 3, p. 252–262, 1 jan. 2005.

SALTON, G.; MCGILL, M. J. **Introduction to modern information retrieval.** New York u.a: McGraw-Hill, 1983.

SARACEVIC, T. An essay on the past and future (?) of information science education—I. **Information Processing & Management**, v. 15, n. 1, p. 1–15, jan. 1979.

STOCKER, M. *et al*. Towards research infrastructures that curate scientific information: A use case in life sciences. In: AUER, S.; VIDAL, M,E. (eds). Data Integration in the Life Sciences. DILS 2018. **Lecture Notes in Computer Science**, v. 11371 p. 15, 2018..

TUDHOPE, D.; BINDING, C. Still Quite Popular After all Those Years— The Continued Relevance of the Information Retrieval Thesaurus. **Knowledge Organization**, v. 43, n. 3, p. 174–179, 2016.

VICKERY, B. C. **Techniques of information retrieval**. London: Butterworth, 1970.

VIEIRA, J. M. L.; SANTOS, M. T.; LAPA, R. C. Estudo da construção e aplicação do tesauro na recuperação da informação de teses e dissertações do programa de pós-graduação em desenvolvimento urbano.In: ENEBD, v. XXXIII, **Anais...** 2010.

Eduardo Ribeiro Felipe

Doutorado em andamento em Gestão e Organização do Conhecimento - Universidade Federal de Minas Gerais.

Mestre em Ciência da Informação, Especialista em Engenharia de Software e Graduado em Tecnologia e Processamento de Dados.

Analista de Tecnologia da Informação na Universidade Federal dos Vales do Jequitinhonha e Mucuri.

Maurício Barcellos Almeida

Information Science, Ph.D.

Pós-doutorado – *State University of New York.*

Pós-doutorado – Faculdade de Medicina, UFMG.

Visiting Researcher Scholar – University of Arkansas for Medical Sciences.

Bolsista de Produtividade – Conselho Nacional de Pesquisa (CNPQ).

Professor Associado - Departamento de Teoria e Gestão da Informação (DTGI) - Universidade Federal de Minas Gerais (UFMG)