UNIVERSIDADE FEDERAL DE MINAS GERAIS
ESCOLA DE CIÊNCIA DA INFORMAÇÃO
PÓS-GRADUAÇÃO EM GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO

Eduardo Ribeiro Felipe

A expansão de *queries* sobre terminologias biomédicas: uma comparação de artefatos de representação do conhecimento para Recuperação de Informações

Belo Horizonte
2020

Eduardo Ribeiro Felipe

A expansão de *queries* sobre terminologias biomédicas: uma comparação de artefatos de representação do conhecimento para Recuperação de Informações

Tese apresentada ao curso de Doutorado do Programa de Pós-Graduação em Gestão e Organização do Conhecimento, da Escola de Ciência da Informação da Universidade Federal de Minas Gerais, como requisito à obtenção do título de Doutor em Ciência da Informação.

Área de concentração:
Ciência da Informação.

Linha de pesquisa:
Arquitetura e Organização do Conhecimento.

Orientador: Prof. Dr. Maurício Barcellos Almeida.

Belo Horizonte
2020

F315e Felipe, Eduardo Ribeiro

A expansão de queries sobre terminologias biomédicas [recurso eletrônico]: uma comparação de artefatos de representação do conhecimento para recuperação de informações / Eduardo Ribeiro Felipe. - 2020.

1 recurso eletrônico (167 f. : il., color): pdf.

Orientador: Maurício Barcellos Almeida.
Tese (Doutorado) – Universidade Federal de Minas Gerais, Escola de Ciência da Informação.
Referências: f. 150-158.
Apêndice: f. 159-167.
Exigências do sistema: Adobe Acrobat Reader.

1. Ciência da Informação – Teses. 2. Representação do conhecimento (Teoria da informação) – Teses. 3.Ontologias (Recuperação da informação) - Teses. I. Título. II. Almeida, Maurício Barcellos. III. Universidade Federal de Minas Gerais, Escola de Ciência da Informação.

CDU:025.4.03

Ficha catalográfica: Maianna Giselle de Paula CRB:2642
Biblioteca Profª Etelvina Lima, Escola de Ciência da Informação da UFMG.

FOLHA DE APROVAÇÃO

A expansão de queries sobre terminologias biomédicas: uma comparação de artefatos de representação do conhecimento para recuperação de informações

EDUARDO RIBEIRO FELIPE

Tese submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO, como requisito para obtenção do grau de Doutor em GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO, área de concentração CIÊNCIA DA INFORMAÇÃO, linha de pesquisa Arquitetura e Organização do Conhecimento.

Aprovada em 27 de agosto de 2020, pela banca constituída pelos membros:

Prof(a). Mauricio Barcellos Almeida (Orientador)
ECI/UFMG [por videoconferência]

Prof(a). Benildes Coura Moreira dos Santos Maculan
ECI/UFMG [por videoconferência]

Prof(a). Daniela Lucas da Silva Lemos
UFES [por videoconferência]

Prof(a). Fabrício Martins Mendonça
UFJF [por videoconferência]

Prof(a). Fernanda Farinelli
IGTI [por videoconferência]

Prof(a). Marcus Vinicius Carvalho Guelpeli
UFVJM [por videoconferência]

Belo Horizonte, 27 de agosto de 2020.

UNIVERSIDADE FEDERAL DE MINAS GERAIS

PROGRAMA DE PÓS-GRADUAÇÃO EM GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO

ATA DA DEFESA DE TESE DO ALUNO
EDUARDO RIBEIRO FELIPE

Realizou-se, no dia 27 de agosto de 2020, às 14:00 horas, Videoconferência, da Universidade Federal de Minas Gerais, a defesa de tese, intitulada *A expansão de queries sobre terminologias biomédicas: uma comparação de artefatos de representação do conhecimento para recuperação de informações*, apresentada por EDUARDO RIBEIRO FELIPE, número de registro 2016662098, graduado no curso de TECNÓLOGO EM INFORMÁTICA, como requisito parcial para a obtenção do grau de Doutor em GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO, à seguinte Comissão Examinadora: Prof(a). Mauricio Barcellos Almeida - ECI/UFMG [por videoconferência] (Orientador), Prof(a). Benildes Coura Moreira dos Santos Maculan - ECI/UFMG [por videoconferência], Prof(a). Daniela Lucas da Silva Lemos – UFES [por videoconferência], Prof(a). Fabrício Martins Mendonça - UFJF [por videoconferência], Prof(a). Fernanda Farinelli - IGTI [por videoconferência], Prof(a). Marcus Vinicius Carvalho Guelpeli - UFVJM [por videoconferência].

A Comissão considerou a tese:

(X) Aprovada

( ) Reprovada

Finalizados os trabalhos, lavrei a presente ata que, lida e aprovada, vai assinada por mim e pelos membros da Comissão.
Belo Horizonte, 27 de agosto de 2020.

Prof(a). Mauricio Barcellos Almeida

Prof(a). Benildes Coura Moreira dos Santos Maculan

Prof(a). Daniela Lucas da Silva Lemos

Prof(a). Fabrício Martins Mendonça

Prof(a). Fernanda Farinelli

Prof(a). Marcus Vinicius Carvalho Guelpeli

A Deus, autor da Vida. A meus pais, Antônio e Julieta. Minha irmã Renata, e minha esposa Luciana.

AGRADECIMENTOS

Ao meu orientador, professor Dr. Maurício Barcellos Almeida. Um agradecimento especial pela sua atenção e dedicação para que este trabalho fosse realizado, algo que foi além do profissionalismo habitual em sua conduta.

À professora Dra. Gercina Lima, pelo apoio no início do programa e pela nossa trajetória passada.

Aos professores(as) doutores(as) membros da banca, Benildes Maculan, Daniela Lucas, Fabrício Mendonça, Fernanda Farinelli e Marcus Guelpeli.

Ao amigo Renato Camargos, pelo companheirismo e apoio.

À amiga Maria José Baños Moreno, pelo apoio neste trabalho, além de outras redações que participamos juntos.

À colega Amanda Damasceno, por compartilhar sua visão sempre esclarecida da Ciência da Informação.

Aos colegas do grupo de pesquisa ReCOL - UFMG, pelas experiências e projetos compartilhados.

Aos amigos da UFVJM, pelo apoio em meio ao trabalho e desenvolvimento desta tese – Patrick Ribeiro, Clayton Brás e o "evangelista Python" Nilo Alexandre. A dúvida entre desenvolver o experimento em Java ou Python acabou quando ele perguntou: "Que código é esse que você está escrevendo?".

RESUMO

A expansão de *queries*, ou consultas, é uma técnica que permite ampliar a capacidade de representação da consulta original, adicionando termos relacionados, de forma a incrementar a correspondência sintática entre o documento e a consulta. A técnica pode ser aplicada em vocabulários controlados de todos os tipos. A presente tese se utiliza de terminologias clínicas para estudar as possibilidades de expansão de *queries* na Recuperação da Informação (RI) de artigos científicos. O objetivo geral é investigar a revocação de artigos científicos no processo de recuperação da informação utilizando dois artefatos de representação da área médica: SNOMED CT e MeSH. Ainda que certas terminologias possam pertencer ao mesmo domínio do conhecimento, suas estruturas correspondentes são organizadas em diferentes modelos. Enquanto a MeSH utiliza estruturas tradicionais de Organização do Conhecimento, no sentido de sua origem na Biblioteconomia; a SNOMED CT utiliza constructos formais, a saber, axiomas ontológicos para definir termos e relações. Embora muito da prática e da literatura atual apontem a RI baseada em técnicas estatísticas como a melhor solução, há também indicações que justificam o uso de terminologias especializadas. Essa percepção influenciou o presente trabalho na direção de evidenciar tais possibilidades a partir de um estudo de caso para comparar duas terminologias da área médica, na recuperação de artigos científicos. Algumas questões preliminares envolviam pensar se o uso de uma terminologia poderia ampliar a revocação de documentos, ou o quão diferente seria a aplicação de diferentes terminologias do mesmo domínio no mesmo conjunto de dados. Para responder a essas e outras questões, foi desenvolvido um *software* para aplicar *queries* e coletar os resultados qualitativos dos dois vocabulários já mencionados. Do ponto de vista da metodologia, o trabalho aborda, através de um estudo de caso, a captação e a estruturação de terminologias biomédicas, a aquisição e o pré-processamento de artigos científicos médicos, bem como a concepção de um algoritmo capaz realizar *queries* submetidas a partir de termos comuns em ambas terminologias. Em termos de resultados, os achados apontam maior revocação para a terminologia MeSH, onde a análise comparativa permitiu inferir princípios importantes como: a) a quantidade de palavras por termo, b) a representação sintática e c) as possibilidades de estruturação terminológica, como principais influências fim de sugerir boas práticas - no contexto da RI - para a comunidade científica que desenvolve e mantém tais artefatos. Como contribuições adicionais, além do *software* desenvolvido, as discussões são relevantes para a Ciência da Informação (CI), em um contexto onde a publicação de artigos científicos vem aumentando significativamente, e as terminologias - artefatos desenvolvidos na CI - podem proporcionar um modelo diferenciado na recuperação da informação.

Palavras-chave: Recuperação da Informação; Artefatos terminológicos; Expansão de query; Correspondência textual.

ABSTRACT

The expansion of queries is a technique that allows to expand the representation capacity of the original query, adding related terms, in order to increase a syntactic correspondence between the document and the query. The technique can be applied to controlled vocabularies of all types. This thesis uses clinical terminology to study the possibilities of expanding queries in the Information Retrieval (IR) of scientific articles. The general objective is to prove a comparison between knowledge representation artifacts for information retrieval. Although certain terminologies may belong to the same domain of knowledge, their features are organized in different models. While a MeSH uses traditional Knowledge Organization structures, in the sense of its origin in Librarianship; SNOMED CT uses formal constructs, namely, ontological axioms to define terms and relationships. However, much of current practice and literature points to IR based on statistical techniques as the best solution, there are also indications that justify the use of specialized terminology. This perception influenced the present work in the direction of evidencing such possibilities from a case study to compare two medical terminologies, in the retrieval of scientific articles. Some preliminary questions involved thinking about whether the use of terminology could extend document recall, or how different the application of different terminologies from the same domain to the same data could be set. To answer these and other questions, a software was built to apply queries and collect the qualitative results from the two vocabularies already mentioned. From the point of view of methodology, the work addresses, through a case study, the capture and structuring of biomedical terminologies, the acquisition and pre-processing of medical scientific articles, as well as the design of an algorithm capable of performing submitted queries from common terms in both terminologies. In terms of results, the findings point to a greater recall for the MeSH terminology, where the comparative analysis allowed to infer important principles such as: a) the number of words per term, b) the syntactic representation and c) the possibilities of terminological structuring, as main influences in order to suggest good practices - in the context of IR - for the scientific community that develops and maintains such artifacts. As additional contributions, beyond the software developed, the discussions are relevant to Information Science (IS), in a context where the publication of scientific articles has increased significantly, and the terminologies - artifacts developed at IS - can provide a differentiated model in information retrieval.

Keywords: Information Retrieval; Terminological artifacts; Query expansion; String match

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ANSI -	*American National Standards Institute*
ASCII -	*American Standard Code for Information Interchange*
BBS -	*Bulletin board system*
BD -	Banco de Dados
CAP -	*College of American Pathologists*
CC -	Ciência da Computação
CD -	*Compact Disc*
CI -	Ciência da Informação
CID -	Classificação Internacional de Doenças
COBOL -	*Common Business-Oriented Language*
CTV3 -	*Clinical Terms Version 3*
DeCS -	Descritores em Ciências da Saúde
DVD -	*Digital Versatile Disc*
EHR -	*Electronic Health Record*
FORTRAN -	*FORmula TRANslation*
FSN -	*Fully Specified Name*
FTS -	*Full Text Search*
FTP -	*File Transfer Protocol*
HTML -	*HyperText Markup Language*
HTTP -	*Hypertext Transfer Protocol*
IA -	Inteligência Artificial
IDF -	*Inverse Document Frequency*
IHTSDO -	*International Health Terminology Standards Development Organization*
ISKO -	*International Society for Knowledge Organization*
ISO -	*International Organization for Standardization*
JSON -	*JavaScript Object Notation*
KOS -	*Knowledge Organization System*
LPO -	Lógica de Primeira Ordem
MARC -	*Machine Readable Cataloging*
MEDLARS -	*Medical Literature Analysis and Retrieval System*
MEDLINE -	*Medical Literature Analysis and Retrieval System Online*
MeSH -	*Medical Subject Headings*
MLDS -	*Member Licensing & Distribution Service*

NIST - *National Institute of Standards and Technology*

NLM - *National Library of Medicine*

NoSQL - Não baseado em SQL

OCLC - *Online Computer Library Center*

OPAC - *Online Public Access Catalog*

OWL - *Web Ontology Language*

PBI - Processo de Busca por Informação

PDF - *Portable Document Format*

PED - Processamento eletrônico de dados

PLN - Processamento de Linguagem Natural

PUBMED - *Free search engine for MEDLINE database*

RF2 - *Release Format 2*

RDF - *Resource Description Framework*

RDFS - *Resource Description Framework Schema*

RI - Recuperação da Informação

SGBD - Sistema Gerenciador de Banco de Dados

SI - Sistemas de Informação

SQL - *Structured Query Language*

SNOMED CT - *Systematized Nomenclature of Medicine - Clinical Terms*

SNOP - *Systematized Nomenclature of Pathology*

SOC - Sistemas de Organização do Conhecimento

SRAM Sistema de recuperação de artigos médicos

SRI - Sistemas de Recuperação de Informação

TF - *Term Frequency*

TI - Tecnologia da Informação

TREC - *Text REtrieval Conference*

TSV - *Tab-separated Values*

UMLS - *Unified Medical Language System*

URL - *Uniform Resource Locator*

UTF - *Unicode Transformation Format*

W3C - *World Wide Web Consortium*

WEB - Internet

WWW - *World Wide Web*

XML - *eXtensible Markup Language*

qSUMÁRIO

1 INTRODUÇÃO

A representação e o registro da informação são um marco na história humana. A necessidade de registrar e transmitir seu conhecimento pode ser evidenciada pela construção da linguagem, pela evolução dos símbolos em sua manifestação pelos pictogramas até chegarmos ao alfabeto moderno. O ser humano percebeu cedo que, ao permitir que nossa espécie fizesse proveito deste mecanismo de comunicação, um grande salto no desenvolvimento pessoal e, consequentemente, coletivo seria possível, pois na medida que mais informação era fornecida, mais conhecimento era gerado. Máquinas, ferramentas e objetos das mais diversas utilidades foram criadas com base no crescimento informacional passado por gerações, e hoje, mais do que em qualquer outro momento na história, a informação transformou-se em pilar social dada sua capacidade de proporcionar o desenvolvimento e transformação nas mais diversas camadas sociais.

A representação informacional passou por muitos estágios acima mencionados até chegar onde a humanidade situa-se agora, em um ambiente em que prevalece a representação digital suportada por dispositivos eletrônicos; onde mais do nunca não é suficiente produzir e armazenar a informação, mas também permitir sua recuperação com precisão. O advento da computação trouxe uma abordagem realmente inovadora. Agora, não apenas o registro e a recuperação da informação podem ser realizados por meio deste suporte eletrônico, mas o horizonte aponta para mecanismos que podem automatizar processos, tarefas, de forma comparável ao raciocínio humano. A tecnologia computacional passa a ir muito além da automação de processos repetitivos e adentra o espaço do que nos diferencia das demais espécies: o raciocínio.

Neste contexto, a abordagem da Ciência da Informação (CI) não poderia ser diferente de sua vocação multidisciplinar. Trabalhar a informação neste âmbito tecnológico exige uma grande aproximação da área de Tecnologia da Informação (TI) (SARACEVIC, 1996) e não apenas a tecnologia como uma grande área, mas em sua especificidade: a Inteligência Artificial (IA). O campo da IA surge fortemente no pós Segunda Guerra Mundial, influenciada pela explosão informacional decorrente do conflito militar (RUSSELL; NORVIG; DAVIS, 2010). O trabalho de Vannevar Bush (BUSH, 1945) sobre o *Memex* continua um grande referencial na CI. Bush foi um dos pioneiros a alinhar o problema informacional com mecanismos capazes de lidar com a organização e recuperação da informação. Em seu artigo ele descreve um dispositivo denominado Memex, capaz de armazenar livros, registros e

comunicações com funcionalidades de recuperação que excedem a velocidade e flexibilidade humanas.

Em 1950 Alan Turing propõe um teste o qual denominou "*imitation game"* (TURING, 1950). Seu artigo define de forma satisfatória uma conceituação de inteligência e inicia com o questionamento: "*as máquinas podem pensar"*? Seu teste é descrito como uma experiência onde uma pessoa em um ambiente isolado faz perguntas para outra "pessoa" em outra sala isolada, na verdade um computador. Se o questionador entender que uma pessoa tem respondido o teste, e não um computador, a máquina então é aprovada, o jogo de imitação obteve sucesso. Esta proposta de experimento é reconhecida e respeitada, ainda hoje, mais de 50 anos depois de Turing ter escrito seu artigo.

Entretanto, outros pesquisadores (RUSSELL; NORVIG; DAVIS, 2010) fazem um contraponto e comparam o processo de desenvolvimento tecnológico à medida em que os avanços sobre a aviação obtiveram sucesso no momento em que engenheiros estudaram a aerodinâmica e não a pesquisa para reprodução de máquinas que pudesse imitar até outros pássaros. Ou seja, a importância da informação com maior potencial declarativo a fim de permitir uma *compreensão computacional* capaz de fazer com que computadores possam expressar sentido em seu processo de tratamento e recuperação (BERNERS-LEE et al., 2006) foi um ponto fundamental para o investimento em pesquisas que abordem inteligência artificial e tratamento informacional.

E quais os avanços da CI e recuperação informacional para os nossos dias? Uma resposta a este tipo de questionamento deve ser observada sobretudo em uma nova forma de abordar a CI: das novas técnicas advindas com o avanço das chamadas tecnologias semânticas, no aprimoramento de artefatos terminológicos (TUDHOPE; BINDING, 2016; HJØRLAND, 2016), nos modelos de representação do conhecimento (ALMEIDA, 2013) e na necessária automação de processos (SALTON, 1996), a CI encontra sua vocação.

Este trabalho reconhece os avanços de décadas de investimento na construção de modelos de organização do conhecimento, especificamente, artefatos terminológicos na área de conhecimento médico. Objetiva-se posicionar estes artefatos em um cenário moderno, onde a recuperação da informação demanda esforços para exercer sua função diante de ambientes e usuários específicos.

De fato, as tecnologias desenvolvidas com o advento da Internet permeiam nosso cotidiano e influenciam as mais diversas iniciativas em tecnologia da informação. E este é um contexto que Ciência da Informação não pode ignorar: a informação em seus mais diversos aspectos, depende de um grande conjunto tecnológico neste período em que vivemos. Os

modelos de indexação, classificação e recuperação que resultaram em grandes avanços no passado, precisam ser revisitados em um cenário eletrônico, dotado de possibilidades e desafios inexistentes anteriormente. Não é apenas uma reflexão sobre uma possível adaptação da CI no contexto digital, acredita-se que é necessário repensar a CI neste novo ambiente.

Nesse contexto, acredita-se que o presente trabalho possui **justificativa** que se estende a diversos aspectos, em particular, à i) necessidade de comunicação científica e ii) a recuperação da informação no meio digital.

Do ponto de vista da comunicação científica, observa-se que a publicação de artigos vem aumentando significativamente no ambiente acadêmico. Este tem sido o meio mais utilizado para comunicação científica e recebe cada vez mais destaque em modelos de reconhecimento tanto na academia quanto em iniciativas do mercado de trabalho. Stocker *et al*. (2018), por exemplo, chegam a afirmar que a publicação de artigos é a única forma de comunicação científica. Outros, como Björk, Roos e Lauri (2008) e Jinha (2010), corroboram com a ideia de que os artigos possuem papel de suma importância na comunicação científica. Por outro lado, Priem (2013) acredita em outros mecanismos além do artigo para a divulgação científica e defende que repositórios também realizam um papel importante no ambiente digital. O fato é que todos os autores reconhecem a importância do artigo enquanto forma de expressão científica e seu relevante impacto na comunidade mundial. Do ponto de vista da recuperação da informação (RI), o assunto é foco de grande investimento nas Ciências relacionadas à informação, pois é peça chave no modelo informacional digital. A capacidade de encontrar a informação está ligada diretamente ao avanço nas técnicas de RI.

Desta forma, o presente trabalho busca investigar, no contexto da Ciência da Informação, como os artefatos terminológicos biomédicos contribuem para que o processo de recuperação da informação seja efetivado, de forma independente dos processos de indexação ou categorização. E para testar as possibilidades de RI em um comparativo entre duas terminologias da mesma área de conhecimento, é verificado através de análises qualitativas, a identificação estatística de recuperação de artigos científicos.

Uma vez justificado o trabalho, esclarece-se o **problema** investigado. O interesse inicial pela temática ocorreu pela constatação de que, artefatos de representação informacionais tipificados como KOS, a exemplo dos Tesauros e Ontologias, possuem diferentes amplitudes terminológicas e conceituais, ainda que ambos sejam capazes de fornecer grande expressividade. Ao considerar tal perspectiva, coloca-se como **pergunta** de pesquisa: *como esses artefatos de representação informacionais, tipificados como KOS se diferenciam na RI a partir da possibilidade de expansão de consultas na pesquisa por artigos científicos?*

Estes artefatos (KOS) possuem como característica principal as relações entre seus termos. Estas conexões permitem contextualizar conceitos e minimizar ambiguidades, constituindo poderosas ferramentas terminológicas que podem atuar em diversos contextos, e que nesta proposta exercem papel principal na tradução terminológica, mitigando problemas linguísticos relacionados à ambiguidade e sinonímia. **Justifica-se**, portanto, investimentos em estudos de modelos e iniciativas que façam o uso de SOCs no ambiente de tratamento e recuperação da informação.

A presente pesquisa, tem como **objetivo geral** investigar a revocação de artigos científicos no processo de recuperação da informação utilizando dois artefatos de representação da área médica, a saber: as terminologias SNOMED CT, muitas vezes considerada uma ontologia (EL-SAPPAGH *et al.*, 2018) e MeSH denominada como um vocabulário controlado, tesauro[1]. Os **objetivos específicos**, por sua vez, são: a) a criação de banco de dados para persistência de artigos científicos para o teste; b) o desenvolvimento de um *software* para a realização de consultas expandidas a partir de relações hierárquicas e ontológicas no banco de dados; c) a submissão das consultas ao banco de dados, bem como sua análise; d) a comparação de resultados de revocação dos dois artefatos.

Acredita-se como principais **fatores motivadores** da proposta do modelo a ser apresentado neste trabalho para recuperação da informação, seja a riqueza conceitual contida nos artefatos terminológicos. Embora cientes de seus desafios devido à imprecisão da terminologia médica, Almeida e Aganette (2017) explicam que testes preliminares indicaram possibilidades reais para uma representação mais rica com o uso de terminologias em comparação com outras técnicas, a exemplo da tradução automática da requisição utilizando dicionários.

O restante do presente trabalho está dividido em duas partes principais: **a primeira parte** contém os capítulos 2, 3 e 4 e traz a parte teórica sobre recuperação da informação e terminologias, suas características e os potenciais desafios para a CI. **A segunda parte** contém a metodologia de pesquisa, a qual descreve as etapas necessárias para alcançar o objetivo, além dos resultados obtidos no desenvolvimento do *software*, comparações terminológicas, discussão e considerações finais.

[1] https://www.ncbi.nlm.nih.gov/mesh/

Parte - I

2 RECUPERAÇÃO DA INFORMAÇÃO

Este capítulo procura apresentar conceitos importantes sobre informação que vão permear todo o trabalho e conduzir diversos processos descritos adiante. A seção 2.1 apresenta uma base sobre Sistemas de Informação e suas origens. A seção 2.2 discorre sobre uma visão geral da Recuperação da Informação. A seção 2.2.1 exibe as origens da Recuperação da Informação e seus principais autores, reconhecendo sua evolução em processo histórico. A seção 2.3 apresenta os conceitos básicos da Recuperação da Informação.

2.1 Sistemas de Informação - Aspectos Introdutórios

Um Sistema de Informação geralmente é um sistema baseado em computador que fornece informação a uma instituição para ajudar a orientar suas ações. O termo “sistema de informação” também é às vezes usado em Ciência da Informação para se referir a Sistemas de Recuperação de Informação centrados mais em documentos do que em dados, domínio de aplicação familiar às bibliotecas e centros de documentação. Às vezes, o termo também é usado de maneira geral e informal, sem referência a computadores ou organizações como por exemplo quando, as pessoas se referem aos seus próprios sistemas de informação pessoais. Nesta unidade, adota-se a perspectiva organizacional, mas que se aplica a instituições de todos os tipos, incluindo bibliotecas. Em uma organização, um Sistema de Informação geralmente se caracteriza por ter pessoas que trabalham interativamente com computadores em tarefas específicas. A interação humano-computador permite que as pessoas e suas extensões em máquinas sejam informadas por meio do sistema, permitindo que decisões rotineiras e altamente estruturadas possam ser muitas vezes automatizadas e delegadas à máquina.

A informação fornecida por sistemas de informação serve para coordenar esforços especializados e coletivos de usuários. Existe uma grande variedade de Sistemas de Informação, o que reflete a diversidade de organizações e de tarefas a serem realizadas. Uma grande instituição pública ou privada possui Sistemas de Informação para dar suporte ao fluxo de trabalho e a funções de contabilidade e finanças, operações, gerenciamento da cadeia de suprimentos, vendas e marketing, atendimento ao cliente, recursos humanos, pesquisa e desenvolvimento. Porém, Sistemas de Informação são encontrados em todos os lugares, em

instituições de todos os tipos e tamanhos, tanto públicas quanto privados (SWANSON; LARSON, 2012).

Segundo Manning, Raghavan e Schütze (2008) a Recuperação de Informação é a procura por material (geralmente documentos) de natureza não estruturada (geralmente texto) que satisfaz a necessidade de informação em grandes coleções (geralmente armazenada em computadores). A RI envolve o armazenamento, organização e pesquisa de coleções. Constitui-se em parte significativa do desenvolvimento tecnológico humano desde a escrita, ou ainda antes. Os primeiros sistemas de RI legítimos foram os esquemas de organização de arquivos e bibliotecas, como os arquivos sumérios ou os "Pinakes[2]" desenvolvidos por Calímaco[3] para a Biblioteca de Alexandria[4]. No século XX, o maior impulso ao desenvolvimento de sistemas automatizados de RI foi a necessidade de gerenciar quantidades cada vez maiores de informação em instituições negócios e no desenvolvimento científico. As primeiras tentativas de automatizar os recursos de pesquisa para coleções de documentos envolveram técnicas baseadas em cartões perfurados, bem como máquinas que utilizavam sensores ópticos em documentos microfilmados (LARSON, 2012).

De acordo com Souza (2006), os sistemas de recuperação da informação organizam e viabilizam o acesso a itens de informação desempenhando as atividades de:

a) Representação das informações contidas nos documentos, usualmente através dos processos de indexação e descrição dos documentos;
b) Armazenamento e gestão física e/ou lógica desses documentos e de suas representações;
c) Recuperação das informações representadas e dos próprios documentos armazenados, de forma a satisfazer as necessidades de informação dos usuários. Para isso é necessário que haja uma interface na qual os usuários possam descrever suas necessidades e questões, e através da qual possam também examinar os documentos atinentes recuperados e/ou suas representações.

O objetivo de qualquer sistema de RI é selecionar itens informativos – textos, imagens, vídeos etc. – aos quais se faz referência como "documentos", relevantes para um determinado pesquisador ou usuário de uma grande coleção de itens. Essas coleções variam de pequenos conjuntos de itens em computadores pessoais aos vastos recursos da Web. Em todos os casos, a tarefa é a mesma: extrair um conjunto de itens que o usuário deseja ter, separando-

[2] https://www.britannica.com/topic/Pinakes
[3] Calímaco (310 a.C. — 240 a.C.), poeta, bibliotecário, gramático e mitógrafo grego
[4] Um dos maiores centros de produção do conhecimento na Antiguidade, estabelecida durante o século III a.C. no complexo palaciano da cidade de Alexandria, no Reino Ptolemaico do Antigo Egito

o daqueles que não desejam. Não se trata de tarefa simples, pois envolve não apenas os aspectos técnicos de desenvolvimento de sistemas, mas também aspectos comportamentais e psicológicos do usuário para entender o que diferencia itens desejados dos não desejados. Na computação é considerada uma área ampla, focada primariamente em prover aos usuários um fácil acesso às informações de seu interesse (BAEZA-YATES; RIBEIRO-NETO, 2011).

2.1.1 Origens dos Sistemas de Informação

Os Sistemas de Informação (SI) modernos surgiram com a disseminação da computação digital na década de 1950, ainda que cartões perfurados estivessem sendo usados para processamento de dados antes disso. O próprio programa armazenado no computador foi inicialmente visto como um dispositivo de cálculo, adequado para análises numéricas sofisticadas. Tal "computação científica" era distinta do que foi denominado "processamento eletrônico de dados" (PED), que surgiu na mesma época para apoiar o trabalho em organizações, como por exemplo na contabilidade.

Na década de 1960, os computadores passaram a ser projetados e comercializados para fins comerciais, finalmente substituindo o equipamento para cartões. Também foi desenvolvida uma linguagem de programação de alto nível para negócios, a *Common Business-Oriented Language* (COBOL), que enfatizava as estruturas de dados e arquivos, além dos recursos computacionais do *FORmula TRANslation* (FORTRAN), a linguagem até então mais utilizada em computação científica. O COBOL finalmente se tornou a linguagem de programação mais usada para o desenvolvimento de aplicativos para SI em computadores de grande porte, os mainframes. Grande parte desse código permanece em uso, e por isso essas linguagens persistem até hoje.

Ainda segundo Swanson e Larson (2012), além do *software* orientado para organizações, o surgimento da tecnologia de banco de dados no final da década de 1960 foi relevante para o aumento e disseminação de SI. Um banco de dados é uma coleção organizada de arquivos de dados relacionados (ELMASRI; NAVATHE, 2011). Um sistema gerenciador de banco de dados (SGDB) é um *software* que permite a gestão de bancos de dados como conjuntos de dados integrados, onde os relacionamentos entre as entidades são delineados. Com um SGDB, os dados podem ser definidos em um dicionário de dados e gerenciados separadamente dos diferentes sistemas que os acessam. Neste processo evolutivo, a articulação do Modelo de Dados Relacional como "padrão de fato" estimulou o desenvolvimento de banco

de dados relacionais na década de 1970, que passou a dominar a área de sistemas até os dias de hoje.

2.2 Recuperação da Informação: uma visão geral

A Recuperação da Informação está intimamente ligada aos processos de representação, armazenamento, organização e acesso a recursos de informação. O conjunto de itens que podem ser processados 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 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 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), Manning, Raghavan e Shutze (2008), Baeza Yates e Ribeiro Neto (2011), dentre outros.

A necessidade do usuário em revisitar a informação é o que norteia os princípios da RI, e esta necessidade é baseada em contextos. Entende-se por contexto as conexões semânticas do usuário em dado momento, visto que um termo de forma isolada pode conter diversos sentidos e seu significado pode ser afetado na utilização da linguagem com outros termos.

Quando *Tim Berners Lee* projetou a *World Wide Web* pensava-se na representação do conhecimento de forma mais ampla (LEE, 2006) e não apenas nas conexões realizadas pelos *hiperlinks* característicos da Internet. Este nível de representação denominada de Web Semântica permitiria que interpretação por agentes computacionais (algoritmos) e possibilitaria respostas a consultas, e não apenas a recuperação em um nível estático, entre o simbolismo do texto de entrada com os documentos que correspondem àquele símbolo.

Para que este modelo de representação fosse efetivamente usado, um conjunto de tecnologias foi criada, o *eXtensible Markup Language* (XML) e o *Resource Description Framework* (RDF). A primeira permite que qualquer pessoa possa criar suas próprias *tags*, marcadores padronizados por caracteres "<" e ">" que permitem adicionar metadados a uma parte da informação, delimitando e estruturando o documento a fim de possibilitar um

processamento passível de interpretação. Mas somente a marcação por meio da XML não consegue determinar vínculos que permitam a construção de estruturas lógicas do conhecimento humano.

A tecnologia desenvolvida em seguida, a RDF, foi a primeira resposta à necessidade de expressividade para efetividade da Web Semântica e sua estrutura compreende em três componentes básicos: *subject* (assunto)*, predicate* (predicado) *and object* (objeto)*,* formando o que se convencionou chamar de tripla (POWERS, 2003). A RDF permite definições simples, mas que não eram possíveis por meio de outras tecnologias. Uma tripla é um fato, uma afirmação (*assertion*); ela pode ser conectada a outras triplas, formando uma corrente de conexões e ampliando um modelo de representação do conhecimento. Deve-se considerar, portanto que apenas as possibilidades criadas pelas triplas em RDF não são suficientes para estabelecer uma descrição do próprio modelo.

O RDF estabelece conexões semânticas por meio da tripla, mas não há uma definição sobre o que cada elemento (assunto, predicado e objeto) é na composição da estrutura do domínio. Para suprir esta necessidade, foi criada o RDF Schema (RDFS). Esta tecnologia é definida pela W3C[5] como: *"um vocabulário de modelagem de dados para dados em RDF".* Powers (2003) contribui fornecendo a definição de que: *"esta tecnologia fornece recursos necessários para descrever os objetos e propriedades de um esquema específico de domínio - é um vocabulário usado para descrever objetos, seus atributos e relacionamentos dentro de uma área específica de interesse".*

RDF e RDFS foram grandes avanços na representação do conhecimento, mas para que as ontologias fossem possíveis, como raciocinadores lógicos (*reasoners*) capazes de descobrir de novas relações e classes, são necessárias outras funcionalidades não presentes em RDF e RDFS. Lacy (2005) enumera algumas limitações da RDFS:

- Há poucos mecanismos de restrição de propriedades em RDFS, por exemplo, não há restrição na propriedade de cardinalidade. E restrição de propriedades são úteis na especificação de membros de uma determinada classe.
- O RDF provê poucos descritores que suportam inferência e regras adicionais são necessárias para que os *reasoners* possam inferir novos fatos.
- O RDFS não provê expressividade suficiente para o nível de descrição que suporte uma ontologia requerida na Semântica Web.

[5] Disponível em: https://www.w3.org/TR/rdf-schema/

A fim de permitir maior expressividade e o suporte a inferência por *reasoners*, foi criada a *Web Ontology Language* (OWL), outra tecnologia criada pelo *World Wide Web Consortium* (W3C) dentro da proposta da Web Semântica. OWL é simplesmente uma linguagem lógica, um subconjunto da Lógica de Primeira Ordem (LPO). Cabe ressaltar que nem tudo que é escrito em OWL é uma ontologia, e nem toda ontologia é ou pode ser representada em OWL (RECTOR; SOTTARA, 2014).

Este período foi marcado pela necessidade de que computadores pudessem "compreender" a informação da mesma forma que os humanos eram capazes de processar e realizar inferências sobre o conteúdo acessado (BERNERS-LEE; HENDLER; LASSILA, 2006). As ontologias podem ser representadas pelas notações RDF e OWL, por meio de um formalismo estruturado, o qual define classes, atributos e, principalmente, relações entre estas entidades, permite que modelos matemáticos sejam aplicados a fim de produzir inferências lógicas. Tal processo não é viável em outros artefatos como as taxonomias e tesauros. E este acréscimo de formalismo é uma das principais características que definem um artefato como ontologia ou não.

Este trabalho compartilha da visão de Almeida (2013) ao seguir a notação de Guarino e Giaretta (1995) onde "ontologia" com letra inicial minúscula, refere-se a um artefato e "Ontologia" com letra inicial maiúscula, refere-se a abordagem filosófica. E embora tanto a CI como a CC façam uso destas conceituações para direcionar suas pesquisas, existem evidências que os sistemas usados em Ciência da Computação e a Ciência da Informação têm objetivos diferentes.

Há três variações da OWL (MOTIK, 2009):

- OWL-Lite: uma versão restrita, destinada a aplicações de banco de dados;
- OWL-DL: versão destinada a ser expressiva tanto quanto possível e ainda permitir que *reasoners* sejam computacionalmente tratáveis;
- OWL-Full: a versão mais expressiva da linguagem onde o suporte aos *reasoners* é incrementado, ainda que em muitos casos seja intratável.

É neste contexto que os *reasoners* atuam, utilizando OWL-DL e algoritmos para realizar inferências sobre a estrutura criada por pessoas, a fim de adicionar à ontologia novas relações e novas classes. Este mecanismo é particularmente relevante porque estas definições através de axiomas, são usadas também em modelos de representação do conhecimento além das ontologias, a exemplo de terminologias clínicas, a fim de processar estruturas hierárquicas dinâmicas, a exemplo da terminologia clínica SNOMED CT. Outra terminologia relevante neste trabalho, a MeSH, usa o formato XML para publicação de sua estrutura terminológica. O

formato, inicialmente projetado em função das demandas no ambiente web, permite publicar as informações para serem consumidas por modelos desktop ou web.

O trabalho em 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. Neste trabalho aceita-se a conceituação de Salton (1972), onde:

i. Revocação - É a proporção de material relevante realmente recuperado;
ii. Precisão - É a proporção de material recuperado realmente relevante.

A formalização de revocação em Baeza-Yates e Ribeiro-Neto (2011) usada neste trabalho está expressa da seguinte forma no quadro 1:

Quadro 1 - Formalização da revocação

$$Revocação = \frac{\{documentos\ relevantes\} \cap \{documentos\ recuperados\}}{\{documentos\ relevantes\}}$$

Fonte: Baeza-Yates e Ribeiro-Neto (2011)

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 CI construiu um arcabouço significativo de conhecimento e 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, como explica Borko (1968), para quem a CI investiga as propriedades e o comportamento da informação, as forças que governam o fluxo de informações e os meios de processar informações visando acessibilidade e a usabilidade. Saracevic (1979) entende que o problema abordado na CI é a eficácia na comunicação do conhecimento público. Dessa forma, a RI se encontra dentre as disciplinas do núcleo duro da CI.

2.2.1 Origens da Recuperação da Informação

A Recuperação da Informação já era parte dos problemas da ciência mesmo antes dos primeiros dispositivos computacionais (ou mesmo eletrônicos) começarem a surgir. Na verdade, os dispositivos eletrônicos trouxeram novas possibilidades na solução de problemas já existentes e conhecidos anteriormente, como a caracterização de documentos e sua posterior recuperação em acervos com número de volumes que já inviabilizam a identificação imediata de determinado documento.

No final da década de 1940 já havia um crescente interesse em sistemas de informação capazes de solucionar a demanda por informação em grandes acervos documentais. Embora o modelo de recuperação baseado em artefato terminológico não tenha sido descartado por completo, a exemplo do trabalho de Bernier e Crane (1948), ganha força o viés da procura independência de artefatos no processo de RI.

Ainda na década de 1940, o renomado cientista Vannevar Bush (1945), diretor do escritório de pesquisa científica e desenvolvimento dos Estados Unidos, publica seu artigo seminal "*As we may think*", com destaque para sua proposta de um mecanismo (*hardware*) que conseguiria fazer o armazenamento e recuperação da informação com a percepção de conectividade; de associação dos elementos informacionais ali submetidos imitando o processo de raciocínio humano. Bush também afirmou em seu trabalho que a dificuldade em recuperar registros estava diretamente relacionada, em grande parte, pela artificialidade dos sistemas de indexação. Mas seu dispositivo nunca chegou a ser construído, e veio a se tornar um modelo conceitual de relativo impacto na Ciência da Informação.

Este foi um período marcado pela necessidade de soluções para desafios antigos já conhecidos pela dinâmica mecânica da recuperação da informação em acervos bibliográficos. Importante contextualizar que neste período histórico o *hardware* de memória computacional (volátil e não volátil) era um item caríssimo, portanto a necessidade da representação documental ao utilizar uma linguagem artificial era um caminho justificável.

Lesk (1996) traça o trajeto da RI a partir de um paralelo com sete etapas da vida de um homem, de Shakespeare. O autor reconhece o impacto do trabalho de Bush na década de 1940 e as limitações tecnológicas da época, principalmente em questões de armazenamento e processamento. É o reconhecimento de um trabalho com viés de predição, não apenas questões de infraestrutura técnica, mas também modelos de interface gráfica, foram tema do trabalho de Bush, como as possibilidades de informação personalizada por interface de usuária, prática extremamente comum nos *softwares* hoje em dia.

A década de 1950 foi marcada por grandes autores como Mooers (1950) e Luhn (1953). Este último, que também contribuiu com os primeiros índices *Key Word in Context* (KWIC), estruturas que destacavam e alinhavam, de forma ordenada, termos chave para permitir consultas a partir deste índice. Luhn (1953) desenvolve um trabalho intitulado "Um novo método para gravação e recuperação da informação". Essa corrida pelo desenvolvimento e reconhecimento de um novo e eficiente método de recuperação da informação é também percebida no sistema "Zatocode" de Mooers (1956). Mooers (1950) publica um trabalho que considerava o uso de um artefato terminológico em associação a mecanismos de recuperação da informação. Na visão de Mooers (1956) as terminologias (linguagens documentais) refletiam uma linguagem apropriada no modelo de representação e recuperação, considerando que termos usados na indexação de documentos possuem boa semântica de representação documental. Um pouco mais tarde Mooers (1960) reconsidera sua estratégia inicial e aponta para um modelo de desenvolvimento de linguagem específica para indexação que não dependa de artefatos terminológicos (ROBERTS, 1984). A partir de 1957 em seu artigo "*A statistical approach to Mechanized Encoding and Searching of Literary Information*", Luhn (1957) explicita o termo "Thesaurus" como o tipo específico de artefato terminológico a ser utilizado em sua proposta.

A década de 1960 trouxe avanços como a definição dos conceitos de revocação e precisão, além do desenvolvimento de mecanismos de avaliação de sistemas de recuperação da informação (LESK, 1996). Em 1964 o *Medical Literature Analysis and Retrieval System* (MEDLARS), baseado no *National Library of Medicine's - Index Medicus* - publica o índice de literatura médica mundial. Mais tarde este índice, acrescido de outros índices é digitado como banco de dados (COOL; BELKIN, 2013). Era um período onde os índices, desenvolvidos por métodos manuais ou eletrônicos, eram impressos para distribuição e consulta. Foi também nesta década a percepção da pesquisa por linguagem natural (*free-text searching*) tomou forma.

A possibilidade de recuperar itens por qualquer termo ao invés de um conjunto restrito de palavras chave começa a tomar forma, embora houvesse limitações técnicas para sua implementação, e também resistências a partir de pessoas que levantaram dúvidas a respeito de que a indexação não seria feita da melhor forma, seja pela a escolha das palavras chave, ou na eleição de termos relevantes dado determinado contexto da área de conhecimento. Essa última justificativa se baseava no cenário onde vários sistemas de indexação já usavam vocabulários oficiais, e poderiam não ter correspondência com os termos de sistemas automatizados. De modo que um usuário poderia procurar os documentos pelo termo "A", mas os mesmos estariam indexados pelo termo "D".

O tesauro MeSH recebe reconhecimento como referência na área médica. E questões sobre como consultas por texto livre poderiam obter resultados com maior qualidade começam a ter grande atenção na ciência da computação. No *Cranfield Institute of Technology*, Cyril Cleverdon desenvolve modelos de avaliação de RI e modelos matemáticos de revocação e precisão. Há uma troca (*tradeoff*) entre estas medidas: se um sistema recupera mais documentos, isso aumenta a revocação, mas diminui a precisão (CLEVERDON; MILLS; KEEN 1966). E foi na *Cornell University* que Gerard Salton desenvolve um modelo vetorial para recuperação da informação, reconhecido e usado atualmente, demonstrando que a indexação manual era lenta e cara, ainda que usada em pequenas coleções.

A indexação por texto livre era realmente efetiva, com custo apenas computacional e disponível para grandes conjuntos de dados, foi um enorme avanço para a área. Também houve esforços para a atribuição automática de termos em categorias de tesauros, de forma automática. Ou ainda, se os tesauros poderiam ser melhor construídos. Uma série de métodos foi discutida em Cleverdon (1970). Ainda nesta década a ideia da relevância por feedback de usuário foi tratada e Salton (1968) ainda contribui com pesquisas sobre recuperação da informação em ambientes bilíngues, usando tesauros e mapeamento de termos em ambos idiomas de uma mesma área. Esforços para o desenvolvimento de Inteligência Artificial e o processamento de linguagem natural marcaram este período (LESK, 1996). Ainda neste momento histórico, a IA contribuiu com o interesse na sumarização automática, técnica que permite a extração de conteúdo capaz de transmitir a ideia principal do texto de forma automatizada, visando lidar com grandes coleções (GUELPELI, 2012).

A década de 1970 inaugura um período de grandes volumes de dados adequados para o processamento computacional, enquanto que na década anterior os primeiros textos começaram a ser digitados e produzidos. Avanços computacionais eram notórios, e demandas onde era necessário o processamento em agendamento (*batch* ou em lote), agora começam a produzir resultados em tempo real. Inicia-se também um período de declínio do custo de *hardware* de armazenamento, o que se sucede ao longo dos próximos anos, continua inclusive agora, no momento da escrita deste trabalho. Reconhece-se que tarefas de processamento textual que demorariam muito tempo para serem processadas, podem ser feitas em fração do tempo manual.

Ainda na década de 1970, a *Online Computer Library Center* (OCLC[6]), uma cooperativa global de bibliotecas com o objetivo de prover serviços de tecnologia, pesquisa e

[6] Disponível na Internet em: https://www.oclc.org

programas comunitários, publicou dados da *Library of Congress* em formato MARC[7] para centenas de bibliotecas membro, fomentando interoperabilidade e trabalho computacional em larga escala. Neste período avançam pesquisas relacionadas a técnicas probabilísticas em recuperação da informação, envolvendo a frequência de palavras em documentos relevantes e irrelevantes, a fim de identificar medidas de frequência dos termos e ajustar o peso a ser atribuído a cada palavra. Na IA o foco é direcionado aos campos de tradução automática, linguística computacional e reconhecimento de voz para sistemas especialistas.

Nos anos 80 o processamento textual é ampliado de forma rápida. Em paralelo acontece o declínio dos preços de *hardware* de armazenamento. Isso impulsiona o processamento em texto pleno (*texto completo*) e não apenas por resumos e palavras-chave. O lançamento dos *CD-ROMs*[8] como nova possibilidade de armazenamento, em contraste com os frágeis e limitados disquetes, abre espaço para distribuição de informações que antes eram limitadas. E mesmo com uma internet de conexão discada, sem a infraestrutura que é comum hoje, aumenta o número de bancos de dados disponíveis pela rede mundial de computadores. Neste período também cresce o número de *Online Public Access Catalog* (OPAC), ou seja, bancos de dados mantidos por bibliotecas ou grupo de bibliotecários (COOL; BELKIN, 2013). A OCLC agora, com mais informações convertidas em formato computacional, disponibiliza consultas *online* e impulsionam a conversão de catálogos para o ambiente digital. Essas mudanças naturalmente demandaram as consultas em texto pleno em larga escala, e não apenas em textos técnicos e científicos, revistas e jornais estavam publicando suas informações *online*, aumentam as pesquisas em novos métodos para RI. Era momento avançar com soluções relacionadas a desambiguidade usando dicionários eletrônicos e terminologias.

Na década de 1990, a infraestrutura em telecomunicações melhorava o tráfego da internet, e o aumento de publicações *online* já era uma realidade em diversos segmentos da sociedade (acadêmico, mídia jornalística, etc). Já haviam iniciativas de comércio eletrônico, mas este modelo não era popularizado ainda. Foi importante perceber que nesta década, muitas pessoas não apenas consumiam informação pela web, mas também começaram a produzir e publicar informação. O protocolo HTTP revoluciona o modo como se lida com informação na rede, o protocolo *Gopher* – um sistema hierárquico de navegação entre servidores – é rapidamente suplantado, da mesma forma que serviços como os *Bulletin board system* (BBS) –

[7] Machine Readable Cataloging

[8] Compact Disc Read Only Memory ou Disco Compacto de Memória Somente Leitura

servidores que permitiam a distribuição de *softwares*, informação, jogos *online* e bate papo (chats).

Além disso, a criação do primeiro navegador (*browser*) com possibilidades gráficas, o *Mosaic*, confirma a internet como modelo padrão de publicação e consumo informacional. *Hardwares* de digitalização como scanner apresentam novas possibilidades de representação da informação e desafios para recuperação, visto que as técnicas baseadas em texto puro (*match-string*) não são válidas para recuperação de imagens. As *Text REtrieval Conference* (TREC[9]), entidade patrocinada pelo *National Institute of Standards and Technology* (NIST) e pelo *U.S. Department of Defense*, encoraja pesquisadores em RI com uma série de oficinas especializadas e fornecendo grandes coleções textuais para modelos de teste e prototipação.

Os anos 2000 marcam a era multimídia. Computadores domésticos com maior poder de processamento, *hardware* de armazenamento em contínuo declínio de preços e aumento de capacidade. Houve o surgimento de mídias como CD e DVD, para leitura e escrita. Foi nesta década que três ex-funcionários do serviço de transações monetárias *online PayPal* criaram o *YouTube*, o maior site de publicação (*streaming*) de vídeos até hoje. Neste cenário a RI explora com intensidade os metadados, informações textuais a respeito de outra informação, textual ou não. Padrões como *Dublin Core,* criado na metade da década passada, a partir de outro padrão da Ciência da Informação: o MARC, promove um caminho de interoperabilidade entre sistemas.

Atualmente, enquanto este trabalho é escrito, o cenário de RI é direcionado em diversos segmentos: sejam os arquivos multimídia e padrões de metadados, as coleções específicas de determinada área do conhecimento, ou ainda a internet e sua miscelânea de padrões e tecnologias. A RI depara-se também com grandes volumes de dados (*big data*) e a necessidade de lidar com este enorme volume de informações para extrair sentido, valor.

Percebe-se, portanto, que mesmo com apenas poucas décadas de pesquisa, a RI mostra-se como uma importante disciplina em uma realidade cada vez mais produtiva, onde informação precisa ser recuperada, com rapidez e precisão.

2.3 Conceitos básicos em Recuperação da Informação

O presente capítulo apresenta uma abordagem introdutória a RI, seus objetivos e propósitos principais. Há um largo histórico desde a antiguidade descrito em Sanderson e Croft

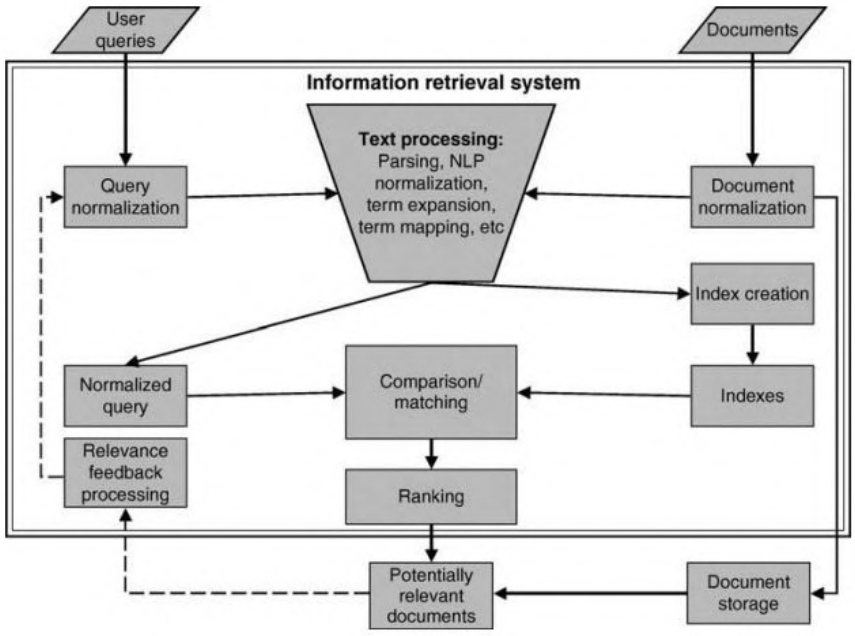
[9] Disponível: https://trec.nist.gov/overview.html

(2012) que detalha cronologicamente o percurso da RI. O propósito deste capítulo é contextualizar a sua importância na temática deste trabalho, apresentando características e apontando desafios motivadores da pesquisa.

Os sistemas de recuperação de informação (SRI) são constituídos por vários componentes de *software* relacionados à suas funções principais, como por exemplo: i) aceitar uma entrada na forma de documentos, extrair informações de documentos e armazená-las em formulários que possam ser acessados para atender à pesquisas do usuário; ii) aceitar consultas do usuário e convertê-las em formulários que possam ser comparados à informações armazenada sobre documentos (LARSON, 2012).

Estes sistemas contam com dois processos principais interdependentes que permite tornar os itens armazenados – sejam registros, documentos ou representações de documentos – acessíveis aos usuários, à saber, indexação e recuperação, conforme a figura 1 abaixo. Embora os SRIs não sejam necessariamente baseados em computador – como no caso de índices impressos e catálogos de cartões – este trabalho se concentra em sistemas baseados em computador.

Figura 1 - Estrutura de um sistema de recuperação da informação



Fonte: Larson, 2012, p. 17

O processo de indexação nos SRI gera uma representação do documento de acordo com regras e procedimentos configurados para o sistema específico. Uma das formas mais simples é extrair todas as palavras que ocorrem em cada documento e armazená-las em um índice, juntamente com uma indicação de qual documento fazem parte. Esse processo é

conhecido como índice invertido, em que determinada palavra está associada a um documento. Como alternativa, um SRI pode usar alguma forma de indexação na qual as palavras são atribuídas a documentos com uma ponderação numérica associada, indicando a importância do significado desse termo no documento.

A indexação também identifica a posição da palavra em um documento, às vezes com a enumeração de parágrafo, sentença ou mesmo a posição da palavra em uma sentença. Essa indexação é usada para determinar restrições de proximidade de uma consulta em um sistema booleano, quando por exemplo as palavras da consulta devem ocorrer na mesma frase ou parágrafo, ou ponderação de proximidade em sistemas de classificação. De fato, a normalização do documento e a criação de um índice envolvem o processamento mais complexo do conteúdo do documento do que a simples extração de palavras.

Neste contexto, utiliza-se "palavras-chave" para indicar elementos de um documento presentes em um índice. Os termos podem ser palavras, frases ou algum tipo de mapeamento do conteúdo do documento usado em vocabulários controlados. Segundo Larson (2012), os estágios típicos do processamento de texto para indexação de documentos podem incluir:

i. Identificação estrutural: reconhecimento e seleção de elementos do documento com base em sua estrutura do documento, por exemplo, seleção do título ou autor; as linguagens de marcação como HTML e XML são usadas em vários documentos disponíveis on-line. Esta codificação permite identificar e extrair os elementos estruturais significativos de documentos.
ii. Tokenização: "*tokens*" são partes do texto, em geral sequências de caracteres alfabéticos e numéricos, extraídos do conteúdo; normalmente, ignora-se o espaçamento, os limites da página e a pontuação e o resultado é o conjunto de *tokens* que podem se tornar termo no índice ou passar para fase de processamento adicional; quando se extraem palavras individuais, as posições originais no conteúdo também podem ser mantidas nos índices, para que, por exemplo, palavras próximas umas das outras no texto original possam ser identificadas.
iii. Reconhecimento de padrões de *tokens*: os *tokens* são examinados para identificar padrões significativos e úteis, o que pode incluir, por exemplo, o reconhecimento de URLs ou identificação de letras maiúsculas sequenciais como potenciais acrônimos.
iv. Normalização de *tokens*: envolve alterar a capitalização de *tokens* para uma única forma (por exemplo, todas minúsculas), remover os diacríticos (acentos, tremados, etc.) e substituí-los por letras, etc.

v. Marcação de partes do discurso: técnicas de processamento de linguagem natural podem ser usadas para identificar as diferentes partes do discurso no conteúdo, ou seja, substantivos e frases substantivas, dentre outros.
vi. Processamento de *stopwords*: os SRIs podem ter listas de palavras inúteis para recuperação, as quais, geralmente, incluem artigos como “a”, “o”, “um”; alguns sistemas possuem listas de *stopwords* diferentes para cada um dos elementos estruturais ou para cada índice; as listas de *stopwords* também podem especificar classes de *tokens* (como números) ou outras partes do discurso.
vii. Análise morfológica ou *stemming*: Muitos SRIs usam alguma forma de derivação morfológica para encontrar várias formas flexionadas de uma palavra, como por exemplo, gatos e gatinhos. O *stemming* usa processamento baseado em regras para executar essa tarefa e nem sempre funciona bem; a análise morfológica se vale de uma combinação de pesquisa e regras de exceção de dicionário, buscando uma única raiz usada nos índices para todas as formas da palavra.

Os SRIs podem incluir combinações e variantes desses processos, por exemplo, alguns sistemas incluem tanto a forma “derivada” quanto a formas original do termo, gerando uma tabela de formas originais para cada ramo encontrado durante a indexação e, em seguida, usando todas as variações durante a pesquisa.

A indexação de documentos e o processamento de consultas exigem procedimentos semelhantes ou etapas de processamento de texto quase idênticas, de forma a garantir que os *tokens* da consulta dos usuários sejam comparáveis aos *tokens* extraídos durante a indexação.

Os resultados do processamento de texto dos documentos são armazenados em índices do sistema, os quais fornecem um mecanismo de pesquisa rápida para cada *token* extraído pelo processamento, juntamente com informações como um identificador para o(s) documento(s) de origem do *token*, informações estatísticas sobre a frequência de ocorrência do *token* em cada documento e na coleção, além do local de origem de cada *token* em cada documento.

As estruturas de arquivos para os índices variam de sistema para sistema, mas as mais comuns estão na forma de índice invertido, em que os dados de todas as ocorrências de um *token* em uma coleção são organizados por uma única instância do próprio *token*. Isso permite que os dados para resolver a maioria das consultas são pesquisados uma única vez no índice para cada *token*. O processo denominado “comparação / correspondência” é responsável por esta tarefa, que é bastante comum em SRIs, independentemente do seu tipo ou modelo. O resultado da correspondência – ou pesquisa de índice – é um conjunto não ordenado de *tokens*

e das informações associadas dos índices que também estavam presentes na consulta normalizada.

2.3.1 Modelos de Recuperação da Informação

A verdadeira diferenciação entre SRIs ocorre nas variações em processamento e classificação de texto. As abordagens para derivar uma lista classificada de documentos relevantes a partir dos resultados da correspondência com índice são chamados de “modelo de RI". Nesta seção, examinam-se os principais modelos de RI e como eles obtêm a classificação de documentos potencialmente relevantes.

De acordo com Larson (2012), existem três classes principais de modelos de recuperação: i) booleano, ii) vetorial e iii) probabilístico. Além disso, existem sistemas híbridos, por exemplo, um sistema vetorial com recursos de limitação de resultados booleanos.

2.3.1.1 Modelo Booleano

O modelo de recuperação mais antigo é o modelo booleano que recebe esse nome por ser baseado na lógica booleana. Os sistemas que usam o modelo booleano foram usados nos mais antigos serviços de busca comercial, catálogos de bibliotecas on-line e sistemas de busca local em sites individuais.

Trata-se de um modelo orientado a conjuntos, em que conjuntos de documentos são definidos pela presença ou ausência de um termo de índice individual. Usando as definições formais acima, um único vetor *Ij* do arquivo invertido pode ser considerado um conjunto booleano, onde cada documento de peso diferentes de zero no vetor define um membro do conjunto. No processamento de consultas booleanas, os conjuntos que representam os diferentes termos de pesquisa podem ser combinados usando as operações do conjunto booleano: interseção, união e negação.

Segundo Manning, Raghavan e Schütze (2008), o modelo booleano é um modelo de recuperação de informação onde pode-se submeter qualquer consulta no formato de uma expressão com termos booleanos, onde os termos são combinados com os operadores: AND, OR e NOT.

a. AND: a interseção gera um novo conjunto de saída a partir de dois conjuntos de entrada em que os membros – ou seja, os documentos compartilhados pelos dois conjuntos – se tornam os membros do conjunto de saída;

b. OR: a união gera um novo conjunto de saída a partir de dois conjuntos de entrada em que os membros de um (ou de ambos) dos conjuntos de entrada são combinados para criar o novo conjunto de saída;
c. NOT: a negação opera em um único conjunto de entrada e gera um novo conjunto que contém todos os membros (documentos) que não estão no conjunto de entrada.

Todas essas operações podem ser executadas usando um arquivo invertido simples ou arquivos invertidos com pesos, mais complexos. Observe-se, no entanto, que a operação NOT pode ser muito cara ao processar, pois o conjunto de resultados seria todos os documentos no sistema não incluídos no conjunto.

Nos SRIs booleanos, o usuário é responsável pela formulação de uma combinação de termos e operadores booleanos na expectativa de selecionar os documentos relevantes para suas necessidades de informação. Os sistemas booleanos geralmente são estendidos pela inclusão de outros operadores, fornecendo restrições adicionais às operações de interseção. Por exemplo, as operações de proximidade de termos restringem a interseção aos documentos que não apenas possuem os dois termos, mas também aqueles a uma distância especificada um do outro (LARSEN, 2012).

Os SRIs booleanos não têm ordenamento padrão para o resultado do processamento de consultas. Qualquer ordem apresentada à pessoa é aplicada após a obtenção do conjunto de resultados booleanos final. Muitos sistemas usam alguma ordem do conjunto de resultados, técnicas rudimentares, podem usar informações do documento, como por exemplo, a data ou o nome do autor. Modelos de ordenamento mais elaborados podem contar com a Frequência do Termo no Documento (*Term Frequency - TF*) onde a frequência que o termo ocorre é armazenada no índice, permitindo uma ordenação por ordem decrescente na coleção resultante. Com um pouco mais de sofisticação, é possível incluir a Frequência Inversa dos Documentos (*Inverse Document Frequency - IDF*) em um modelo denominado TF-IDF, onde a presença dos termos nos documentos frente à coleção, é levado em conta no cálculo. Caso o termo esteja presente em muitos documentos, sua importância será penalizada e vice-versa (BAEZA-YATES; RIBEIRO-NETO, 2011).

2.3.1.2 Modelo Vetorial

O modelo de espaço vetorial considera cada vetor de documento como descrito sobre um vetor em um espaço dimensional “M”, isto é, uma dimensão para cada termo na coleção. Os SRIs de espaço vetorial baseiam seu ranqueamento na proximidade entre os vetores

de documentos e o vetor de consulta no âmbito do espaço M. Isso corresponde ao cálculo de uma medida de similaridade baseada em termos usados na consulta e nos documentos da coleção. Na prática, o arquivo invertido é usado para garantir que o processamento considere apenas documentos com pelo menos um termo em comum com a consulta. Caso contrário, em coleções típicas, muita computação seria necessária para processar todos os vetores de documentos, mesmo quando eles não mantivessem termos comuns com a consulta (LARSEN, 2012).

Na RI, as técnicas de correspondência parcial baseadas no espaço vetorial modelam documentos na coleção em ordem decrescente de similaridade com a consulta. Os termos da consulta são ponderados por frequência de termos simples, e o peso do documento é calculado como o produto interno dos vetores de consulta e de documento.

Como principais vantagens do modelo vetorial, pode-se destacar:

1. Seu esquema de peso por termo permite maior qualidade na recuperação.
2. Em estratégias de comparação parcial, é possível recuperar documentos que estão mais próximos da descrição da query inicial.
3. Sua fórmula de ranking baseada em cosseno ordena os documentos de acordo com seus níveis de similaridade com a query.
4. A normalização da extensão do documento é naturalmente incorporada ao ranking.

Segundo Baeza-Yates e Ribeiro-Neto (2011), teoricamente o modelo tem como desvantagem que os índices de termos são mutualmente independentes. Embora na prática considerar a dependência de termos não é algo simples, e pode levar a resultados insatisfatórios se não for feito de maneira apropriada.

2.3.1.3 Modelo Probabilístico

No modelo vetorial, a suposição subjacente é que documentos semelhantes a uma consulta são relevantes para necessidades de informações do usuário que faz tal consulta. A relevância é geralmente definida como uma avaliação subjetiva por parte de certo usuário sobre o valor ou a utilidade de um documento para satisfazer uma necessidade. Pode-se dizer que, para uma pesquisa em um SRI, cada um dos documentos da coleção contém informações que o indivíduo deseja. Se isso acontecer, o documento é considerado relevante, caso contrário não. O problema na recuperação de documentos é selecionar todos os registros do repositório que o usuário considera relevantes e rejeita aqueles que seriam considerados não relevantes.

Dada a natureza individual dos julgamentos de relevância, assume-se que não há relações puramente determinísticas entre termos usados na indexação e a relevância de um documento; ou mesmo entre os termos usados na busca de um documento e a relevância desse documento. Na verdade, essas relações são de natureza probabilística: não há soluções perfeitas para a recuperação de documentos, mas apenas soluções aproximadas. De acordo com este modelo e o trabalho teórico em RI, a abordagem apropriada é classificar cada documento em ordem decrescente de sua probabilidade de relevância para um certo usuário e consulta. Essa declaração é conhecida como princípio de ranqueamento por probabilidade (LARSON, 2012).

O modelo probabilístico foi proposto em 1976 por Robertson e Jones. Dada uma query pelo usuário, há um conjunto de documentos que contém exatamente os documentos relevantes e outro conjunto não. Considera-se o conjunto relevante como ideal, dada uma descrição deste conjunto ideal de respostas, pode-se recuperar os documentos relevantes. Pode-se pensar que o processo de construção da *query* é um processo que especifica as propriedades ideais para o conjunto de resposta. O problema é que não se sabe como são exatamente essas propriedades. O que se sabe é que existem índices de termos, cuja semântica pode ser usada para caracterizar estas propriedades. Desde que estas propriedades não são conhecidas em tempo real da *query*, deve-se realizar um esforço para inicialmente prever quais seriam. O palpite inicial permite a geração de uma descrição probabilística preliminar de um conjunto ideal de respostas, que poderiam ser usadas para recuperar um conjunto de documentos.

Para aprimorar a descrição probabilística de um conjunto ideal de respostas, uma interação com usuário é iniciada. O usuário pode observar e avaliar nos documentos recuperados da primeira iteração, quais são relevantes e quais não são. O sistema então pode usar essa informação para refinar a descrição do conjunto ideal para resposta. Ao repetir este processo algumas vezes, é esperado que esta descrição da *query* ideal seja aprimorada e tornando-se cada vez mais precisa (BAEZA-YATES; RIBEIRO-NETO, 2011).

2.3.1.4 Outros modelos

Ainda segundo Larson (2012), outros modelos de importância são os modelos de linguagem, os modelos de redes de inferência e os modelos híbridos.

A modelagem de linguagem foi originalmente desenvolvida para aplicativos de reconhecimento automático de fala, tradução automática e correção de *Optical Character Recognition* (OCR). Um modelo de linguagem para SRIs é simplesmente a distribuição estatística dos termos nos documentos e na coleção como um todo. No SRI baseado em modelo

de linguagem, em vez de estimar a probabilidade da relevância, o sistema tenta estimar a probabilidade de uma consulta específica ter sido gerada a partir do modelo de linguagem para certo determinado documento. Os resultados são ranqueados de acordo com essa probabilidade. A suposição é que se a consulta pode vir do documento, é provável que tal documento seja relevante.

Os modelos de rede de inferência são uma forma de modelo probabilístico que usa redes de probabilidades conhecidas de eventos – como a probabilidade a priori de um termo em um documento – para inferir a probabilidade de eventos desconhecidos como por exemplo, a probabilidade de relevância para um determinado documento. As redes de inferência podem incluir nós que são resultados de operações booleanas na rede de inferência probabilística. As redes neurais são semelhantes às redes de inferência, onde as probabilidades de relevância, dadas entradas específicas, são estimadas pelo treinamento da rede que envolve o ajuste das probabilidades de ativação de baixo nível de um nó para qualquer outro nó conectado. A principal diferença é que, em uma rede de inferência, as informações inferenciais são pré-codificadas no modelo como estrutura e parâmetros, enquanto nas redes neurais nada é codificado a priori, mas uma tabula rasa em branco é extensivamente treinada com dados e as inferências descobertas são resultado desse treinamento.

Muitos SRIs combinam dois ou mais dos modelos básicos já discutidos em um único sistema: são os modelos híbridos. A combinação mais comum é permitir alguma forma de operação ou restrição booleana em um conjunto de resultados ranqueados. Isso pode envolver o uso de operações booleanas para restringir um conjunto de resultados ranqueados à itens que satisfazem a restrição booleana. Outros sistemas híbridos podem usar métodos de "fusão de dados" para combinar resultados de diferentes algoritmos de recuperação. A combinação de elementos probabilísticos e booleanos, bem como operadores para suportar várias operações de combinação também são utilizados em muitos sistemas.

2.3.2 Busca pela informação

A busca por informação é afetada por diferentes fatores, dos quais quatro tipos principais determinam a seleção e aplicação de diferentes estratégias de busca: a) objetivo e tarefa; b) estrutura de conhecimento do usuário; c) projeto de sistemas de RI; e d) contexto social e organizacional. Neste contexto, as subseções abaixo discorrem sobre três níveis principais: i) táticas e movimentações; ii) estratégias; iii) padrões de uso.

2.3.2.1 Táticas e movimentações

Táticas são movimentos que os usuários aplicam ao processo de busca. Diferentes tipos de táticas desempenham papéis diferentes na assistência a usuários que buscam por informação. Com base em suas funções no processo de busca por informação (PBI), as táticas de informação podem ser classificadas em táticas de monitoramento, táticas de estrutura de arquivos, táticas de formulação de busca e táticas de termos. Segundo Xie (2012), enquanto táticas de monitoramento e de estrutura de arquivos são usadas para rastrear a busca e explorar a estrutura de arquivos de forma a encontrar a informação desejada – seja uma fonte ou um arquivo – táticas de formulação de busca e táticas de termos são aplicadas para auxiliar na formulação e reformulação de buscas, como bem como para ajudar a selecionar e revisar termos. Além dessas táticas de busca, as táticas relacionadas a ideias ajudam os usuários a identificar novas ideias e soluções para problemas que enfrentam na busca por informação. Enquanto as táticas de geração de ideias incluem pensar, debater, meditar, dentre outros; as táticas de quebra de padrões consistem em captura, quebra, violação etc. Com foco no gerenciamento de tópicos, as táticas de busca baseadas no conhecimento são outro tipo que amplia, restringe e altera o escopo do tópico sob busca.

Semelhante às táticas, as movimentações de busca ilustram como os usuários interagem diretamente com os sistemas de RI *online*. Os movimentos de busca em geral estão relacionados à formulação e reformulação de consultas. Eles podem ser classificados com base no fato do significado de uma consulta ter sido alterada. Enquanto movimentos operacionais mantêm o significado dos componentes da consulta permanecem inalterados, os movimentos conceituais alteram o significado desses componentes. Movimentos conceituais estão associados aos resultados da busca e tem por objetivo reduzir ou aumentar o tamanho de um conjunto recuperado ou melhorar a precisão e a recuperação. Os movimentos de busca também podem ser classificados com base no fato do movimento estar relacionado a movimentos conceituais ou físicos. Movimentos cognitivos se referem aqueles que os usuários fazem, conceitualmente, para analisar termos ou documentos. Movimentos físicos, por sua vez, se referem aos movimentos que os usuários fazem para usar os recursos do sistema (XIE, 2012).

2.3.2.2 Estratégias

As estratégias também podem ser definidas em diferentes dimensões de interação dos usuários com sistemas de RI e com objetos de informação incorporados aos sistemas. Busca

e navegação são as principais estratégias que usuários empregam quando interagem com sistemas de RI. As estratégias de navegação exigem mais interações do que as estratégias de busca analítica. Já as estratégias ativas e reativas especificam outra abordagem para classificar estratégias de busca. Ao aplicar estratégias de planejamento, usuários tomam decisões sobre como procurar informação antes dos primeiros movimentos, como por exemplo, autor, título, conceito relacionados, suporte externo, recursos do sistema, etc. Ao aplicar estratégias reativas, os usuários tomam decisões seguindo um movimento após o outro, como mudanças de foco, relacionamentos de termos de busca, recuperação de erros e assim por diante.

Xie (2012) observa que as estratégias de busca em mecanismos de recuperação na Web têm características próprias. Estratégias de busca que se concentram na reformulação de consultas que foram geradas com base na análise de logs (registros de acesso) e são: específica, genérica, paralela, de componentes básicos, dinâmica, multitarefa, recorrente e de reformulação. Algumas são semelhantes às estratégias em banco de dados on-line, como específica, genérica e de componentes básicos, mas outras revelam características exclusivas de estratégias de mecanismo de busca Web, como multitarefa, recorrente, dinâmica, dentre outras. Em ambientes de mecanismos Web, os usuários realizam diferentes tarefas simultaneamente; as buscas são mais dinâmicas e geralmente aplicam as mesmas consultas repetidamente. O ambiente da Web também define projeto e recursos exclusivos da busca. Algumas estratégias de solução de problemas representam as estratégias de busca Web incluem: levantamento, verificação, exploração, acompanhamento de *links*, retrocesso e avanço, busca de atalhos e meta-busca.

2.3.2.3 Padrões de uso

A busca na Web adiciona novo significado à busca em estratégias de busca, em particular, a análise de logs de transações. Diferentemente dos estudos sobre estratégias de busca, os padrões de uso identificados no mecanismo de busca na Web concentram-se em padrões de formulação e reformulação de consultas, com base na análise dos logs. Os padrões de formulação e reformulação de consultas na Web podem ser caracterizados de cinco maneiras: i) consultas curtas; ii) sessões curtas com reformulações mínimas; iii) uso mínimo de operadores e modificadores de busca (nem sempre utilizados corretamente); iv) resultados mínimos de visualização; e v) busca em tópicos (variam de entretenimento, recreação e sexo, comércio eletrônico, etc.). A análise de logs não é apenas quantitativa: facetas de formulações de consulta também são identificadas.

Os padrões de uso em diferentes tipos de ambientes de RI também foram comparados, sendo identificadas semelhanças e diferenças. Consultas curtas, sessões curtas, exibição mínima dos resultados da busca e consultas exclusivas semelhantes foram realizadas na Web e bibliotecas digitais. As sessões de busca variam nos ambientes OPAC, embora os também exibam consultas curtas. Mais consultas contêm operadores booleanos em ambientes de bibliotecas digitais do que em ambientes de mecanismos de busca na Web. O contexto das bibliotecas digitais representa um híbrido de RI "tradicional", usando recursos bibliográficos, e RI "popular" exemplificada por sistemas de busca pública. O uso de mecanismos on-line em bibliotecas digitais revela que os tópicos de busca dos usuários estavam próximos dos bancos de dados das bibliotecas digitais, mas seus comportamentos de busca eram mais semelhantes aos mecanismos da Web. Em geral, os usuários se envolvem em buscas mais extensas em ambientes de RI tradicionais, como OPAC e banco de dados *online* (XIE, 2012).

3 ARTEFATOS TERMINOLÓGICOS

Esse capítulo apresenta os Sistemas de Organização do Conhecimento que influenciaram esta pesquisa e a conceituação de expansão de *queries*. Na seção 3.1.1 os tesauros são apresentados. Na seção 3.1.2, as ontologias.

A RI recebe importância crescente, tanto na CI, como na CC. Está relacionada com a representação, armazenamento, organização e acesso a itens informacionais (SALTON; MCGILL, 1983). É uma área presente nos ambientes físicos e digitais, e que neste trabalho foi tratada no ambiente computacional. Segundo Vickery (1970), o problema da recuperação da informação não é novo pois existem ferramentas para identificar documentos – bibliografias, catálogos, índices – há séculos, mas o termo "recuperação da informação" raramente foi encontrado antes de 1955.

Quando Tim Berners Lee projetou a *World Wide Web* (WWW) pensava-se na representação do conhecimento em uma forma mais ampla (BERNERS-LEE; HENDLER; LASSILA, 2006) e não apenas nas conexões realizadas pelos *hiperlinks*, característica marcante da Internet. Este nível de representação denominada de Web Semântica permitiria que a informação fosse interpretada por agentes computacionais (algoritmos) e possibilitasse respostas de caráter interpretativo, e não apenas a recuperação informacional em um nível estático entre o simbolismo do texto de entrada com os documentos que correspondem àquele símbolo. Neste universo de publicações eletrônicas, a recuperação da informação possui relevantes desafios. Mesmo projetada para ser um ambiente aberto, baseada em hipertexto e independente de formatos proprietários, a Internet não restringe as formas de publicação e exposição de dados. O resultado é um ambiente altamente diversificado em formatos que em grande parte permitem sua recuperação de forma aberta e livre de impedimentos por formatos abertos de codificação.

São inúmeras pesquisas que dependem de dados para análise, sejam eles estruturados ou sem estrutura formal. E "estrutura" compreendida neste contexto, diz respeito à forma como os dados estão organizados, se possui algum padrão para identificação computacional ou não. Um exemplo de dado estruturado pode ser compreendido pela listagem de clientes de uma loja, onde as informações estão delimitadas por agrupamentos bem definidos. Dados não estruturados por sua vez estão relacionados a textos em linguagem natural, que embora possuam estrutura formal do idioma, não possui regras bem definidas para o processamento computacional na extração de informações. Algumas iniciativas foram criadas

para estruturar dados não estruturados no ambiente web, a exemplo da adoção de linguagens de marcação como XML, XHTML. Estruturar dados permite que o algoritmo computacional siga regras claras para identificação de uma informação específica. No mundo “ideal” os dados publicados na Internet seriam estruturados de tal forma que os algoritmos pudessem realizar sua recuperação de forma inequívoca; mas estamos longe desse modelo. Para que a recuperação aconteça precisa-se trabalhar com diversas técnicas a fim de possibilitar a extração de dados na Internet e em ambientes não estruturados.

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 (HOPPE; HUMM; REIBOLD, 2018). Ainda segundo estes autores tal tradução é feita por KOS do tipo 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. Ampliando o campo de recuperação para o usuário na identificação de documentos que não estariam presentes em uma recuperação tradicional. Chowdhury (2004) elenca 7 grandes funções em um sistema de recuperação da informação:

1. Identificar as fontes de informação para as áreas de interesse da comunidade;
2. Analisar o conteúdo das fontes, muitas vezes, documentos;
3. Representar o conteúdo de fontes a fim de adequá-las para as requisições dos usuários;
4. Analisar requisições e representá-las para pesquisa no banco de dados;
5. Coincidir a requisição de pesquisa com a informação no banco de dados;
6. Recuperar a informação que é relevante;
7. Fazer ajustes necessários no sistema baseado em feedbacks de usuários.

A capacidade dos usuários em representar e expressar suas questões a fim de serem processadas em um modelo de RI, também deve receber especial atenção. Pode-se citar cinco grandes classes envolvidas em sistemas de RI:

1. Usuário;
2. Pergunta;
3. Pesquisador;
4. Pesquisa;
5. Itens recuperados.

Assume-se que o contexto da pergunta é composto por um conjunto de variáveis que afetam os eventos de busca e recuperação da informação. Questões são expressões linguísticas, criadas com a intenção de induzir respostas (SARACEVIC, 1988). A capacidade de representar o desejo do usuário em confronto com a representação da linguagem autoral, é o

desafio que este trabalho deseja abordar. Na medida em que esta relação não é atendida em grande parte, qual mecanismo pode mediar essa combinação terminológica, conceitual, a fim de alinhar o resultado com a perspectiva da pergunta.

Há também a atenção com o perfil do usuário no processo de RI. Não apenas seu conhecimento prévio, ou ausência dele, mas também na capacidade de lidar com modelos de recuperação em sistemas computacionais. Em geral pode-se entender que a maioria dos usuários possuem boa adaptação e manuseio nas tarefas computacionais. Os sistemas operacionais modernos estão muito mais amigáveis, iniciativas de IA começam a ser usadas na prática, como reconhecimento de voz, entre outras medidas a fim de facilitar a operação computacional.

Borgman (1989) detectou em seu experimento que, a aptidão técnica e características de personalidade influenciam o processo de operação em sistema de RI. Naquele momento, interfaces booleanas foram melhores usadas por usuários de perfil técnico, enquanto outros tipos de interface foram melhores para usuários não técnicos. O estudo ainda afirma que as características pessoais em sistemas de RI são menos influenciadas por características de personalidade, levando a descobertas de que princípios técnicos eram diferenciais no processo de recuperação da informação.

Pao *et al.* (1993) também realizou uma pesquisa de usuários, estudantes de medicina, a fim de identificar comportamentos em relação ao processo de pesquisa no ambiente MEDLINE. Entre os questionamentos do trabalho, dois chamaram atenção: a) "como estudantes de medicina, a experiência acumulada em pesquisas no MEDLINE poderia incrementar o resultado da consulta?", em outras palavras, a experiência dos estudantes de medicina com o MEDLINE poderia aumentar as chances de recuperação de artigos mais úteis? b) há alguma relação entre a efetividade da consulta e o nível de conhecimento clínico? Ressalvado o contexto do experimento e sua metodologia, as descobertas foram importantes para aquele momento:

- A grande maioria dos alunos que possuía conhecimento prévio foi capaz de completar as pesquisas e 86% tiveram resultados relevantes dos itens recuperados;
- Não foi encontrada relação entre o conhecimento prévio na elaboração de consultas e o uso de recursos disponíveis na interface do sistema;
- A maioria dos usuários optam por usar uma interface de consulta com estratégias simples, mesmo possuindo níveis de experiência diferenciados;
- Identificou-se uma tendência, onde estudantes que fizeram uso de mecanismo de busca com mais frequência, de fato, produziram melhores resultados;

- Houve uma surpresa ao constatar que estatisticamente, não houve diferença significante nas consultas de alunos em diferentes níveis de conhecimento com o MEDLINE e seu desempenho no currículo escolar.

Sistemas de Organização do Conhecimento (SOC) – ou em língua inglesa, *Knowledge Organization System* (KOS) – é uma expressão que categoriza um conjunto de outros artefatos na CI. Segundo Mazzocchi (2019) pode-se considerar como SOC:

- Cabeçalhos de assunto;
- Tesauros;
- Esquemas de classificação;
- Ontologias;
- Etc;

Souza *et al*. (2010) pondera um conjunto mais abrangente em grandes grupos:

- Texto não estruturado (resumos);
- Termo e/ou lista de conceitos (dicionários, glossários, etc...);
- Conceitos e relações estruturais (vocabulário controlado, taxonomias, tesauros, esquemas de classificação, etc...);
- Relações conceituais e estruturas de layout (mapas mentais, modelos de dados, modelos de referência, modelos de entidade-relacionamento, etc...)

Souza, Tudhope e Almeida (2012) ainda alerta para o fato que não importa o quão extenso seria a tentativa de elencar os diversos artefatos considerados como SOCs, essa tentativa é falha devido às diferentes interpretações possíveis a respeito do que pode ou não, ser considerado SOC.

De acordo com Gilchrist (2003), vocabulários controlados (como os tesauros) foram usados para indexação e recuperação, inicialmente empregados em cartões perfurados, mais tarde, em ambiente digital nos computadores. E embora tenham sido grandemente usados, foram sobrepujados relativamente por técnicas de pesquisa em texto livre e pela Internet. O autor ainda considera o surgimento do "tesauro de pesquisa", direcionado à RI e que tanto os tesauros como as taxonomias e ontologias, possuem como ponto em comum, a relação com a linguagem natural.

Em face às ontologias, Almeida, Mendonça e Aganette (2013) entendem que a possível vantagem na adoção de relações formais em relação à RI, está na possibilidade de automatizar a expansão de consultas, conforme outros autores apontaram, mas essa evidência carece de experiências empíricas. Essa abordagem possui ainda um caráter motivador nesta pesquisa, visto que uma terminologia a ser usada no processo de expansão de *queries* possui

um nível de formalismo ontológico na definição de sua estrutura, enquanto a outra terminologia não possui.

O presente trabalho dá destaque a duas terminologias clínicas consideradas como SOCs: SNOMED CT e MeSH.

3.1 Modelos semi-estruturados

Informações semiestruturadas são caracterizadas por uma estrutura sem tipagem rígida, sem um domínio que possa especificar estritamente o tipo de informação. Em alguns casos um padrão estrutural pode ser identificado, em outras situações não há informações descritivas associadas.

3.1.1 Tesauros

Dos diversos artefatos utilizados pela CI, destaca-se neste trabalho os tesauros, cuja etimologia vem do Grego (thesaurós) e indica as definições: armazém / depósito, ou tesouro (VICKERY, 1970). O tesauro enquanto artefato é uma linguagem pós-coordenada, que teve origem como resposta aos unitermos, em linguagem natural, e aos cabeçalhos das listas de cabeçalhos de assuntos, que são pré-coordenados, e também aos sistemas de classificação enumerativos e hierárquicos. Hoje, é um instrumento de estrutura facetada, pois incorporou os princípios da teoria da classificação facetada, que tem por base os conceitos, representados pelos termos-descritores. Os conceitos não são "unidades de pensamento", mas são entendidos como "unidade de conhecimento", pois são representações de um recorte da realidade, dentro de um contexto, e, assim, perdem a sua aderência aos sentidos extralinguísticos, que permitem interpretações diferentes (semiótica). No sentido de um conceito está delimitado e explicitado em uma definição que pode ser intencional: estrutura superordenada/genérica, extensional: tipos de enciclopédica, funcional, etc (MACULAN, 2020[10]). Segundo Currás (1995), é um vocabulário especializado, no qual as palavras que o compõem estão relacionadas umas com as outras semanticamente. E que segundo os manuais da Unesco tem como definição em relação à sua função: "os tesauros constituem um artefato de controle terminológico usado para transferir os descritores retirados da linguagem natural dos documentos para um sistema linguístico". E sobre sua estrutura: "os tesauros são vocabulários controlados e dinâmicos de termos relacionados, semântica e genericamente, que cobrem um domínio específico do

[10] Informação pessoal, passada por e-mail.

conhecimento". O tesauro tem ênfase nos conceitos que representam os objetos e nas relações que estabelecem entre si, são elaborados a partir de necessidades específicas, pois pode haver, dentro de um mesmo domínio, diversos tesauros contemplando, cada um, um aspecto do domínio tratado (LARA, 2001). E ainda, de acordo com Maculan e Aganette (2017), os tesauros são desenvolvidos para representar o conhecimento de um domínio a partir de um conjunto de termos descritores, preferidos e não preferidos, inter-relacionados em um sistema conceitual, fazendo o controle da terminologia em níveis diferenciados de controle e padronização.

Segundo Walker (2001), os tipos de tesauros incluem: a) tesauro artesanal - estruturas hierárquicas de termos relacionados desenvolvidos por humanos; b) dicionário de sinônimo automático - estrutura com termos relacionados, derivados por coocorrência estatística de palavras ou relações lexicais. São ainda considerados sistemas de classificação de assuntos, que tratam de um sistema linguístico no qual os componentes principais são os termos (CURRÁS, 1995). Entende-se ainda como termos a definição de Felber (1984): "Unidades linguísticas de um vocabulário especializado".

Este artefato possui como características predominantes, uma organização em estrutura hierárquica e semântica, 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: Termo geral (broader term)
- NT: Termo específico (narrower term)
- RT: Termo associado (related term)
- SN: Sinônimo (synonym)
- USE: Use ao invés (use instead)
- UF: Usado para (used to mean)

Todavia a construção de um tesauro exige o conhecimento e experiência de um profissional da informação. A escolha dos termos e suas relações é o que define a qualidade e a abrangência de utilização do artefato em sua adoção prática: tanto em processos diferentes como a classificação ou indexação, quanto na recuperação informacional.

Estes artefatos terminológicos são usados na tradução entre a linguagem (terminologia) usada pelo usuário em contraste com a terminologia da base de conhecimento e, no interesse deste trabalho, do vocábulo médico. 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 envolve vocábulos médicos. O processo histórico dos tesauros é bem referenciado no trabalho de Currás (1995) e

Mendes, Reis e Maculan (2015). Além desses, vários autores convergem no entendimento que este artefato 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).

Currás (1995) e Vieira, Santos e Lapa (2010), por exemplo, reforçam 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 os KOS a exemplo dos tesauros, como principais fornecedores de descritores na indexação, onde os termos são convertidos em palavra-chave, pois determinam o assunto de que trata o documento. As relações terminológicas se apresentam na forma associativa ou por semelhanças de equivalência, tratando-se de uma linguagem para fins documentários que pode ser usada nos processos de indexação ou classificação e na recuperação da informação (CURRÁS, 1995).

Os tesauros situam-se como artefatos onde os componentes principais são os termos (léxicos dedicados a um assunto concreto). São ainda considerados uma linguagem especializada, normalizada, pós-coordenada, para fins documentários, visto que os elementos linguísticos que o compõem, termos simples ou compostos, encontram-se relacionados entre si, sintática e semanticamente. Assume-se que os termos selecionados para indexação possuem correspondência semântica na representação do conteúdo, e por isso podem ser usados como base na RI, além de permitir expansão de consultas pela qualidade de suas relações, segundo Bechhofer e Goble (2001). Os tesauros foram concebidos como ferramentas para RI (ROBERTS, 1984; GIUNCHIGLIA; DUTTA; MALTESE, 2014).

É aceito que a classe gramatical que mais está presente nos tesauros são os substantivos, e como características relevantes pode-se identificar: a especificidade, a exaustividade, composição dos termos (unitermos ou compostos), as relações de sinonímia, as relações de hierarquização e de associação. Schütze e Pedersen (1997) entendem que um tesauro é útil para a Recuperação da Informação quando é específico o suficiente para oferecer sinônimos de termos usados na definição do corpus de interesse.

3.1.2 Ontologias

As ontologias, enquanto artefatos informacionais, são consequência da necessidade de representação do conhecimento, fomentada no início da chamada "Web Semântica".

Uma corrida tecnológica teve início a fim de implementar semântica, prover significado às informações, de forma a permitir que sistemas computacionais fossem capazes de extrair conexões com outras bases de dados e informações não estruturadas realizando inferências e permitindo a produção de conteúdo não estático com base nas relações identificadas.

A Ontologia, assim como as teorias da classificação, tem seu referencial pioneiro em Aristóteles. Na obra do filósofo, este termo foi usado para denominar um ramo da metafísica, direcionada ao estudo das categorias, das entidades existentes e de como estão relacionadas (LOWE, 2006). O termo ontologia ainda possui diferentes significados em áreas distintas:

i. Na Ciência da Computação, a ontologia representa um artefato na engenharia de *software* (GRUBER, 1993);
ii. Na Ciência da Informação, a ontologia é vista sob a ótica bibliográfica e multimídia (LEMOS; SOUZA, 2020), como o estudo da representação de assuntos (VICKERY, 1997).

Almeida (2013) afirma que ontologia na Ciência da Computação 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 Ciência da Informação, ontologias são também usadas para a construção de estruturas de categorias na representação de conteúdo documental. A definição da W3C[11] sobre ontologia sob o viés da Ciência da Computação destacado por Yu (2011, p.137) esclarece que:

> Uma ontologia define formalmente um conjunto comum de termos usados para descrever e representar um domínio. […] Uma ontologia define os termos usados para descrever e representar uma área de conhecimento.

Essa conceituação permite abranger um conjunto maior de artefatos informacionais como as taxonomias e tesauros, considerando-os como tipos de ontologias. Gruber (1993) apresentou uma definição que se tornou clássica do conceito ontologia: "Uma ontologia é uma especificação explícita de uma conceituação."

[11] Disponível em: https://www.w3.org/TR/webont-req/

A Ciência da Computação usa a ontologia para categorizar o mundo, mas enfatiza o processo de raciocínio. A ênfase no raciocínio se encaixa no reino das lógicas, abordado com o objetivo de descobrir como um padrão de silogismo combina duas premissas para chegar a uma conclusão. Ao considerar a representação do conhecimento, as ontologias são consideradas estruturas de conceitos representados por um vocabulário lógico. Um elemento que compôs o conjunto de tecnologias rotuladas por "web semântica" (ALMEIDA, 2013). O autor ainda contribui na observância de dois significados para o termo ontologia em CC: O uso de princípios ontológicos para o entendimento e modelagem da realidade, linha esta alinhada com o papel ontológico na Filosofia, de modo a descrever daquilo que existe e caracterizar as entidades nas atividades de modelagem. O outro significado consiste no conjunto de declarações expressas em uma linguagem de representação, que podem ser processados por mecanismos de inferência automatizados (ALMEIDA, 2014).

O trabalho de Silva, Souza e Almeida (2012) relata uma importante base teórica e exposição de métodos na comparação entre a construção de vocabulários controlados e ontologias. Com destaque para a maturidade da construção do tesauro frente às metodologias de construção de ontologias. Alinhado a este tema, é relevante a metodologia descrita em Farinelli *et al*. (2016) e Farinelli *et al.* (2019), que culminou na ontologia de domínio OntONeo[12], na área de conhecimento obstetrícia e neonatal. E ainda, ressalta-se a metodologia *OntoForInfoScience* por Mendonça e Almeida (2016) para a construção de ontologias no domínio sanguíneo. Estes trabalhos são de grande relevância e contribuem no sentido de estabelecer métodos e técnicas capazes de direcionar artefatos ontológicos que possam ser trabalhados em interoperabilidade, além de permitir integrações com algoritmos no desenvolvimento sistemas de RI, na construção do conhecimento.

3.2 Vocabulários médicos

O vocabulário médico envolve um grande conjunto termos específicos, abreviações e por vezes, ambiguidades. Este cenário impõe desafios a representação do conhecimento médica em ambientes computacionais, a fim de permitir a automação de análises, manipulação e recuperação da informação. No passado, momento histórico na computação, período marcado pela escassez de espaço para armazenamento de dados, alto custo de memória e baixo poder de processamento; as informações eram armazenadas em sistemas que basicamente indexavam

[12] Ontology of Obstetric and Neonatal domain

metadados para representação dos documentos e sua consequente recuperação. Os vocabulários controlados exerceram grande influência neste processo, a fim de fornecer termos para o processo de indexação (CIMINO, 1995).

O processo da utilização de linguagens artificiais é direcionado à solução de um problema reconhecido na Ciência da Informação: o distanciamento entre a representação do usuário e a representação do autor do documento, para os mesmos conceitos. E neste sentido as terminologias exercem uma possibilidade especial para a transposição sintática na convergência do mesmo objetivo semântico. Mas o custo relativo ao processo manual de construção destes artefatos é um fator importante. Direcionar especialistas de domínio para a manutenção e construção de terminologias é por vezes, um fator limitador para início e continuidade destes projetos.

Há de se considerar uma evolução importante neste processo histórico: os primeiros mecanismos de indexação, categorização e, consequentemente, os modelos de recuperação, eram baseados apenas em termos selecionados ou resumos para representação documental. O armazenamento digital era de alto custo, inviabilizando grandes repositórios de dados. Com o advento de novos *hardwares*, a produção em larga escala, entre outros fatores, possibilitou que as bases de dados contivessem o documento em sua integralidade. O que afetou diretamente a construção de índices e o processo de representação e recuperação. Avanços relevantes na RI devem ser reconhecidos pelo modelo vetorial e demais trabalhos em Salton (1972).

Alguns trabalhos atribuem o uso com sucesso das terminologias enquanto suporte na RI, em direta intercessão com o perfil do usuário (ALLEN, 1991; LIU; WACHOLDER, 2017). Mas como será visto adiante, este trabalho objetiva desvincular este perfil, permitindo analisar se, de modo independente do nível de conhecimento do usuário, termos significativos possam participar da pesquisa e qual seu impacto na precisão e revocação em consultas ao acervo de artigos.

Importante notar que em 2015 o renomado capítulo da ISKO-UK[13], organizou um debate cuja proposição era: “Esta instituição acredita que não há mais lugar para os tesauros tradicionais em sistemas de recuperação modernos”. Esta afirmação deve ser vista com cautela e não de forma generalizada. Como um tipo de vocabulário controlado, estes artefatos ainda exercem importante contribuição como poderá ser visto adiante neste trabalho. Como impacto desta discussão, Hjørland (2016) publica seu artigo: “Há lugar para o tesauro tradicional nos modernos sistemas de recuperação?” em uma sinalização importante para revisar o

[13] https://www.isko.org/allevents.php?in=Uk

posicionamento destes artefatos terminológicos, face as modernas técnicas de indexação e recuperação da informação, iniciativas lideradas pela Ciência da Computação. Os posicionamentos do autor serão considerados adiante neste trabalho.

O papel das terminologias assemelha-se à função dos tesauros[14] (CLARKE 2016). E seu controle terminológico depende da precisão da linguagem especial e seu propósito em representar uma determinada perspectiva no domínio (HJØRLAND, 2016).

3.2.1 MeSH

Na presente seção descrevem-se histórico, características, e implementação da *Medical Subject Headings* (MeSH). A terminologia MeSH enquanto SOC é designada como um tesauro, um vocabulário controlado e hierarquizado produzido pela *National Library of Medicine* (NLM[15]). É largamente reconhecido como um dos mais sofisticados, e considerado como "estado da arte" em vocabulário controlado para sistemas de recuperação da informação no domínio biomédico (HERSH, 2009; LIU; WACHOLDER, 2017). Possui atualizações com periodicidade diferenciada para cada formato de distribuição, publicadas tanto como consulta *online* pela ferramenta MeSH browser, quanto para download em diversos formatos, como: XML, ASCII, MARC21 e RDF.

3.2.1.1 Histórico

De acordo com o prefácio[16] do MeSH na NLM, muitos sinônimos, termos relacionados e próximos conceitualmente são incluídos no vocabulário como termos de entrada, com a finalidade de ajudar usuários a encontrar o descritor MeSH mais relevante ao conceito procurado. Vários sistemas relevantes fornecem acesso ao MeSH como terminologia de apoio, a exemplo do MeSH *Entrez* - um conjunto de bancos de dados projetados para ajudar quem pesquisa no *MEDLINE / PubMed*. O *UMLS Metathesaurus* que contém conexões com outros vocabulários controlados. Além da própria interface MeSH web, onde pode-se ter acesso a todo o vocabulário de forma *online*.

A primeira lista oficial de cabeçalhos de assunto foi publicada pela NLM em 1954. O trabalho foi baseado na lista de autoridade interna para a publicação da lista atual de literatura

[14] https://www.isko.org/cyclo/thesaurus
[15] Disponível em: https://www.nlm.nih.gov/
[16] Disponível em: https://www.nlm.nih.gov/mesh/intro_preface.html

médica, que por sua vez, incorporou os termos do catálogo de índices da biblioteca e do cabeçalho de assunto do índice médico trimestral de 1940. Com o início do índice médico em 1960, surge o novo e revisado cabeçalho de assuntos médicos (*Medical Subject Headings*).

Mas vários problemas eram nítidos, por exemplo, “uso terapêutico” foi cadastrado sob “agentes físicos”, “drogas” e “produtos químicos”. “Terapia” foi usado juntamente com “doenças”. Subseções foram definidas para minimizar tais problemas, a exemplo de “terapia de”, “uso terapêutico de” e “aspectos terapêuticos”, foram ligados a “terapia”. O que não resolveu problemas relacionados a sobreposição de significado das próprias subposições. De modo que era difícil decidir se um artigo científico sobre “química biossíntese” ficaria em “química” ou “metabolismo”.

As listas categorizadas de termos foram impressas pela primeira vez em 1963, e continham 13 categorias principais e um total de 58 grupos, separados em subcategorias e categorias principais. Este modelo de listas categorizadas permitiu aos usuários encontrar muito mais termos relacionados do que na estrutura de referência cruzada anterior. Em 1963 a segunda edição do MeSH continha 5.700 descritores, em comparação com os 4.400 da edição de 1960. Dos cabeçalhos usados na lista de 1960, 113 foram retirados em favor de termos mais recentes. A efeito de comparação, a edição de 2015 do MeSH contém 27.455 descritores.

Realmente 1960 foi um ano especial, pois a biblioteconomia médica passava por grandes transformações e a primeira edição da série *Index Medicus* foi publicada. O projeto de informatização também se iniciava e a NLM se preparava para armazenar e recuperar informações em formato digital. O sistema de Análise e Recuperação de Literatura Médica (MEDLARS) agilizaria o processo de publicação de bibliografias como o *Index Medicus* e facilitaria a expansão da cobertura de literatura, além de possibilitar pesquisas por usuários sob demanda.

Uma nova lista de tópicos introduzida em 1960 foi a base da operação de análise e recuperação. O MeSH então era uma versão nova e completamente revisada das listas de cabeçalhos de assuntos, compiladas pela NLM. Uma lista única poderia e deveria ser usada para ambos os propósitos: indexar periódicos e catalogar assuntos. Além de permitir duas grandes vantagens: a) simplicidade para os usuários ao facilitar o aprendizado com apenas um único esquema; b) economia para a Biblioteca no desenvolvimento e manutenção de um único esquema.

3.2.1.2 Características

Desde o início, o MeSH foi concebido com o objetivo de ser uma lista dinâmica, dotado de métodos para recomendar e examinar a necessidade de novos termos. O conteúdo do vocabulário relacionou-se ao uso de termos da própria literatura e evoluiu para atender a novos conceitos no campo da medicina. O advento da computação possibilitou revisões de forma mais prática e sistemática, apesar da dificuldade na atualização de índices impressos e catálogos de cartões.

A NLM em seu site[17] publica um treinamento organizado em oito módulos, que permite ter uma noção clara dos principais princípios desta terminologia. Cada módulo é organizado em subseções que organizam a temática por assunto, facilitando o acesso a determinado item específico (ROMANO, 2018).

Vocabulários controlados como o MeSH são ferramentas importantes no acesso a coleções bibliográficas. Liu e Wacholder (2017) ressaltam que discussões sobre o questionamento da utilização destas terminologias foram levantadas no período em que a recuperação da informação avança em suas técnicas de texto livre. Importância reconhecida pelos trabalhos apresentados em Salton (1983) e Spark Jones e Willett (1997).

Autores diretamente ligados à ciência da informação também teceram reflexões a respeito deste tema, como Calhoun (2006) e Hjørland (2016). Discussões a respeito da indexação automática como processo substitutivo da indexação manual, por especialistas e artefatos como as terminologias são debates ainda atuais na ciência. De um lado, elenca-se o custo de especialistas, a demora do processo manual e a possível imprecisão pelo erro humano. Por outro, a imprecisão da classificação / indexação automática por algoritmos que estão em constante aprimoramento, são contrapontos que alimentam este debate. O propósito dos vocabulários controlados está relacionado com os problemas causados pelo uso da linguagem natural na RI, a exemplo de homônimos e sinônimos (SVENONIUS, 1986).

A adoção de vocabulários controlados é direcionada ao aprimoramento da precisão e revocação em sistema de RI. Svenonius (1986) considera que a necessidade de vocabulários controlados foi baseada, primariamente, na suposição de que a linguagem natural não é sistemática o suficiente para representar conceitos complexos de modo previsível, e, portanto, tem um contexto limitado no acesso aos documentos relevantes. Estes artefatos atuam, portanto,

[17] https://www.nlm.nih.gov/tsd/cataloging/trainingcourses/mesh/intro_010.html

para mitigar problemas linguísticos de ambiguidade e mediar a representação entre a linguagem do usuário e autor.

O trabalho de Abdou e Savoy (2008) aborda, com destaque no estudo em RI com efetividade em meta-informação, avanços ao considerar a inclusão de termos do MeSH tanto em artigos científicos, quanto na concepção de *queries*. Por outro lado, Cimino (1995) aponta cautela no uso indiscriminado de termos de vocabulários controlados em sistema de indexação automática, o que pode causar inadequação ou redundância terminológica.

Este modelo de estrutura informacional com utilização de terminologia de domínio, está diretamente ligado ao nível de conhecimento de indivíduos em uma área de conhecimento específica. Liu e Wacholder (2017) compreendem que:

- O conhecimento da área do domínio ou o tópico específico não está correlacionado com o desempenho da pesquisa, mas usuários experientes em um assunto são hábeis para usar vocabulários controlados na expansão de *queries* e obter melhores resultados na pesquisa;
- Experiências com pesquisas em bancos de dados *online*, não podem predizer a performance da consulta, mas a revocação é otimizada com a experiência e usuários experientes podem encontrar documentos relevantes com mais eficiência do que usuários não experientes;
- Há um efeito de interação na intercessão entre o conhecimento do domínio e a experiência da pesquisa.

As autoras em seu trabalho fazem dois importantes questionamentos que parecem muito pertinentes para os objetivos desse trabalho:

i. Linguagens de indexação controladas podem ajudar usuários a produzir melhores resultados de pesquisa?
ii. Linguagens de indexação controladas podem ajudar diferentes tipos de usuários a produzir melhores resultados de pesquisa?

As discussões apontaram para resultados diferentes baseados em perfis de usuários distintos. O uso de termos da terminologia MeSH foi mais útil para usuários experientes, com domínio da área de conhecimento, em relação aos usuários que não possuíam este embasamento. Houve medições estatísticas que demonstraram um relacionamento significante entre o tipo de usuário e o uso da terminologia.

Em geral, usuários com maior conhecimento do domínio (área do assunto) foram mais aptos a obterem um resultado significativamente melhor do que usuários experientes na construção de consultas com apoio de termos MeSH. A pesquisa apontou também que usuários

treinados tiveram melhores resultados que usuários sem treinamento, no cenário em que termos MeSH não eram disponibilizados. Porém, usuários treinados não extraíram benefícios da terminologia, provavelmente porque não tinham o nível de conhecimento de domínio para entender os tópicos para consulta.

Os resultados apresentados na pesquisa revelaram a importância do conhecimento do domínio para um melhor aproveitamento da terminologia MeSH em sistemas de recuperação da informação. As descobertas apontaram que, conhecer a área do assunto é crucial para uma pesquisa técnica, baseada em tópicos, e que usuários neste perfil que usaram a terminologia MeSH, melhoraram significativamente a precisão de suas consultas. Em geral, usuários com conhecimento do domínio, são os que mais se beneficiaram do uso da terminologia. As descobertas do estudo apontaram que usuários com conhecimento de domínio foram capazes de fazer bom uso de vocabulários controlados, em modelos de expansão de *queries* (ALLEN, 1991; LIU; WACHOLDER, 2017).

Já o trabalho de Pao *et al*. (1993), aponta em sua experiência com 184 estudantes que, embora fizessem uso da terminologia MeSH para a construção da consulta, não houve melhoria efetiva no modelo de RI e que usuários com alguma experiência prévia na construção de *queries*, obtiveram melhores resultados. Deve-se observar, entretanto, a diferença entre públicos participantes entre os diferentes autores, suas condições de pesquisa e condições de objetivo na consulta. O artigo evidencia as seguintes afirmações:

- Houve evidência para afirmar a existência da relação entre o nível de conhecimento prévio na construção de consultas e o subsequente uso de artefatos de pesquisa como MEDLINE;
- Foram encontradas relações entre o conhecimento prévio do usuário e o uso de recursos disponíveis pelo sistema ou banco de dados;
- Foi assumido que o nível elevado de experiência do usuário em consultas, produz melhores resultados na pesquisa.

Shultz (2006) explorou questões relacionados ao grande número de acrônimos usados na linguagem médica para seu mapeamento na terminologia MeSH. O estudo explorou as possibilidades de recuperação da informação no ambiente MEDLINE usando acrônimos e iniciais mapeadas para a terminologia, que por sua vez permitiu conexão com outros termos, ampliando o espectro dos itens recuperados.

Nelson, Johnston e Humphreys (2001) realizou um importante trabalho a respeito das relações terminológicas no MeSH. Reconhece que as hierarquias são a chave para permitir recuperações expandidas, e enumera os tipos de relação existentes no instrumento. A

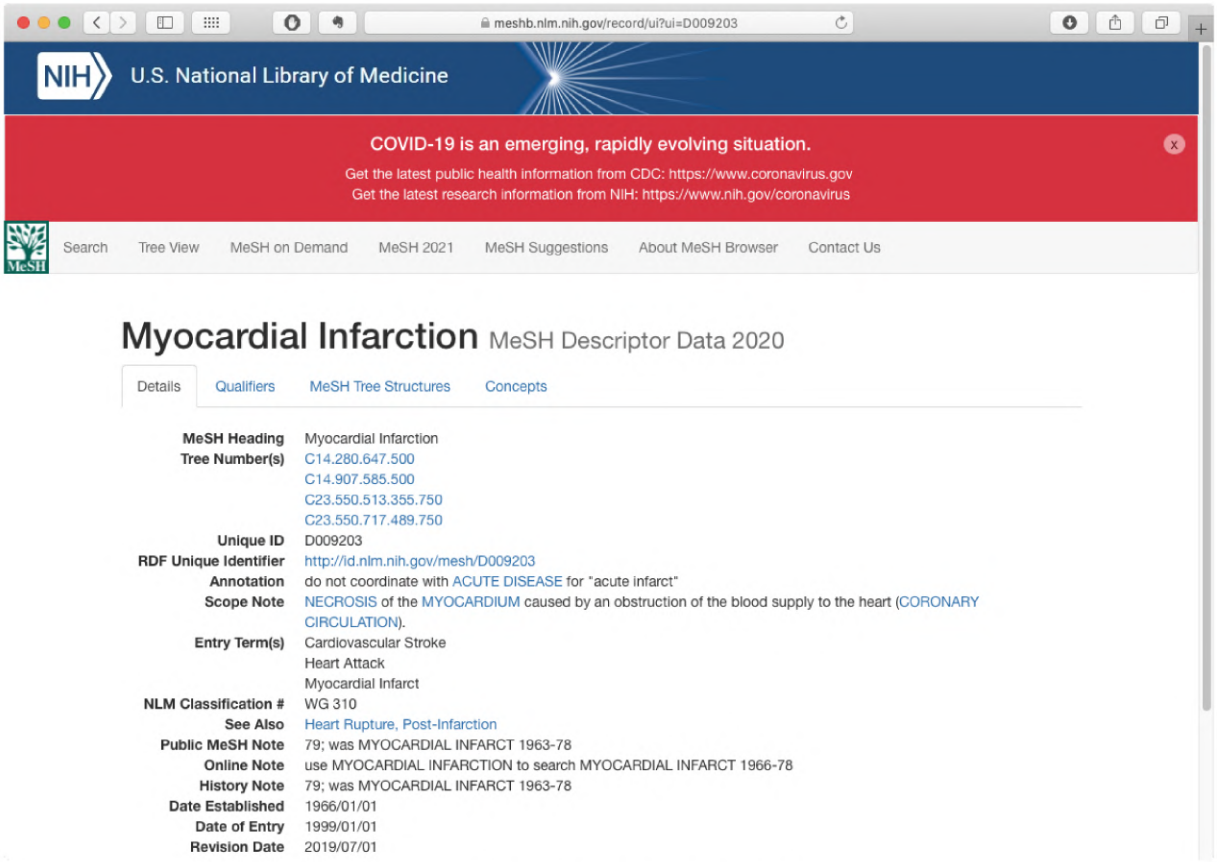
terminologia clínica MeSH é usada amplamente para fins de indexação e catalogação por bibliotecários no mundo todo. Sua estrutura é baseada em três grandes componentes:

i. Os cabeçalhos;
ii. Os subcabeçalhos ou qualificadores;
iii. Registros de conceitos suplementares.

Importante destacar ainda as relações de equivalência, que incluem os termos de entrada e sinônimos, e as relações hierárquicas e associativas. Em seu trabalho, o autor considera que a definição de conceito é a ideia comum, ou o sentido expressado por termos ou palavras sinônimas. E para propósitos computacionais, a identificação de sinônimos está alinhado a característica de todos os termos serem membros de uma mesma classe conceitual. No MeSH, um conceito possui um ou mais descritores. Sendo que na possibilidade de vários descritores, um deles é selecionado como preferido, possuindo um código (*unique identifier*) para possibilitar uma identificação única.

O formato escolhido para manipulação do vocabulário MeSH está codificado em XML. Uma característica importante do MeSH é sua grande extensão terminológica, o que afeta diretamente o tamanho do arquivo eletrônico responsável pela sua representação. Mesmo que sua publicação eletrônica seja dividida em arquivos separados por idioma, cada arquivo em seu tamanho em megabytes é considerável, o que dificulta a manipulação do algoritmo para pesquisa em tempo real, de modo que a terminologia será exportada para o modelo relacional.

Figura 2 - Interface web da terminologia MeSH



Fonte: https://meshb.nlm.nih.gov/search

3.2.2 SNOMED CT

3.2.2.1 Histórico

Em 1955 o *College of American Pathologists* (CAP) iniciou o desenvolvimento da *Nomenclature for Anatomic Pathology*. Em 1965 a CAP criou a *Systematized Nomenclature of Pathology* (SNOP), que mais tarde daria origem ao SNOMED. Inicialmente publicado pelo CAP, o SNOP foi criado para descrever morfologia e anatomia. Em 1975 sob a direção do Dr. Roger Cote, a CAP expandiu o SNOP para criar a *Systematized Nomenclature of Medicine* (SNOMED[18]), a fim de atender as crescentes necessidades da medicina. Em 1979 foi publicado o SNOMED II, tornando-se a versão mais adotada na época. Na década de 80 o projeto *Read Codes* foi desenvolvido de forma independente no Reino Unido pelo Dr. James Read sob a *National Health Service Centre*. Mais tarde este trabalho evoluiu para a versão 3 do *Clinical Terms Version 3* (CTV3). Em 1993 a principal expansão e revisão do SNOMED foi publicada

[18] https://www.nlm.nih.gov/healthit/snomedct/snomed_overview.html

e denominada SNOMED International, ou SNOMED 3.0. Em 1998 a CAP iniciou um projeto de três anos a fim de mesclar as terminologias SNOMED e CTV3. Em 2000 a CAP colaborou com a *Kaiser Permanente* para criar uma versão do SNOMED International chamada SNOMED RT (*Reference Terminology*), publicada inicialmente no ano 2000. Em 2002 um marco importante foi estabelecido, a CAP conclui a fusão das terminologias SNOMED RT e CTV3, e publica a primeira versão do SNOMED CT. Em julho de 2003 a *National Library of Medicine* (NLM) em nome do Departamento de Saúde e Serviços Humanos dos Estados Unidos, firmou acordo com a CAP para disponibilizar o SNOMED CT aos usuários dos EUA, sem custo, através do *Unified Medical Language System* (UMLS) Metathesaurus. O contrato forneceu à NLM uma licença perpétua para o SNOMED CT e suas atualizações contínuas. Em 2007 os direitos de propriedade intelectual de todas as versões do SNOMED foram adquiridos pela recém-criada *International Health Terminology Standards Development Organization* (IHTSDO). Em 2013 o governo dos EUA solicitou que o SNOMED CT fosse incluído no *EHR System*, a fim de serem certificados para uso pleno em estágio 2. Em 2017, todas as versões evolutivas do SNOMED, exceto a SNOMED CT, foram marcadas como fora de uso pela IHTSDO (BHATTACHARYYA, 2016).

A IHTSDO ao adquirir o SNOMED CT, teve como objetivo promover a adoção internacional do SNOMED CT. Sua organização permitiu que a terminologia tivesse estratégias para lidar com novos conteúdos, e um mecanismo de atualização e distribuição que fosse imparcial e transparente para todos. Esta instituição é responsável pela manutenção, desenvolvimento, garantia de qualidade e distribuição contínua do SNOMED CT.

Os Estados Unidos é um dos nove membros fundadores da IHTSDO, juntamente com Austrália, Canadá, Dinamarca, Lituânia, Holanda, Nova Zelândia, Suécia e Reino Unido. Atualmente há a figura de membros (38) e países afiliados. Membros podem ser agências de governo ou corporações, endossadas por autoridade governamental do território que representa.

A necessidade de uma terminologia médica projetada para o ambiente computacional e que pudesse evoluir como padrão multinacional remonta há muitos anos como visto no processo histórico do SNOMED CT. Mas é importante citar outro projeto que teve o mesmo objetivo, o Projeto GALEN[19]. Na década de 1990 este projeto foi criado a fim de representar informações clínicas de uma nova maneira. Era um projeto multilíngue, com abordagem qualitativamente diferente, focando em cinco princípios: a) na interface com o usuário baseada em descrições ao invés de seleção de códigos; b) na estrutura baseada em

[19] http://www.opengalen.org

descrições compostas ao invés de códigos enumerados; c) no estabelecimento de padrões como modelo de referência padrão, em contraste com sistemas de codificação; d) na publicação, como um serviço de terminologia dinâmica de *software*; e) na apresentação multilíngue. O projeto GALEN está inativo e sua última publicação, na versão 8, data de 28 de janeiro de 2010.

3.2.2.2 Características

O *Systematized Nomenclature of Medicine - Clinical Terms* (SNOMED CT) é uma terminologia biomédica muito abrangente e multilíngue. Usada em mais de 80 países conveniados, é distribuída por uma licença denominada *Member Licensing & Distribution Service* (MLDS). Ao realizar a solicitação para acesso à terminologia, o pesquisador recebe um login para acesso ao painel de controle (*dashboard*) onde poderá realizar o download das publicações (*releases*), que acontece duas vezes ao ano.

As raízes desta terminologia estão ligadas ao registro eletrônico de saúde, na identificação de requisitos de usuários (pacientes), à experiência prática e nos princípios científicos estabelecidos em publicações por pares. Seu objetivo, segundo a MLDS é tornar os registros eletrônicos médicos dotados de melhor significado possível, melhorar a comunicação e ser capaz de destacar informações relevantes, permitindo a efetiva recuperação baseada em conceitos. Seu foco é representar informação clínica de forma significante, como parte de um registro eletrônico de saúde bem projetado.

Diferente de outras terminologias, a SNOMED CT trabalha com três pilares na constituição da terminologia: conceitos, descrições e relacionamentos. O conceito é uma representação única, em formato numérico. É um identificador que permite agrupar diversas descrições e relacionamentos de forma não ambígua, oferecendo grande flexibilidade à terminologia. Entende-se ainda que o conceito é uma ideia clínica, representada por um identificador. A este identificador está associado um nome especificado principal (*fully specified name* - FSN), que não pode ser alterado. Este FSN é uma frase que de maneira não ambígua descreve o conceito e possui uma *tag* entre parênteses para expressar a hierarquia daquele conceito.

A descrição é usada para representar cada conceito e há dois tipos de descrições: a FSN e os *Synonym* (sinônimos). As FSN são descrições preferenciais para representação do conceito. São importantes principalmente quando há diferentes conceitos referenciados na mesma representação sintática. Cada conceito pode ter apenas uma descrição FSN, mas ao mesmo tempo possuir vários sinônimos. Essa flexibilidade aumenta a expressividade da

ferramenta, incluindo termos capazes de referenciar um único conceito, com diversas representações. Esta estrutura de representação é a mesma nas suas diversas traduções, nos idiomas já publicados deste instrumento.

Cada conceito pode possuir várias descrições. Estas descrições são termos associados a um ou vários conceitos para expressar aos humanos a representação daquele conceito. A documentação da terminologia afirma que todos os conceitos podem ter extensões adicionais para definições textuais que ajudam a descrever melhor o conceito. Visto que há uma multiplicidade de termos associados a um mesmo conceito, a terminologia permite a escolha de um termo preferido (*preferred term*), capaz de ser selecionado como a descrição preferencial em um cenário de múltiplas escolhas.

Um grande diferencial desta terminologia está em permitir a conexão de um conceito para outro conceito, de forma lógica, usando axiomas. Estas conexões podem ser poli hierárquicas, proporcionando uma grande flexibilidade na definição da terminologia. E são as estas possibilidades de relações, uma das principais características que tornam essa terminologia tão dinâmica e expressiva. Cada conceito pode se relacionar a outros conceitos por meio de relações. Pode-se definir diversas relações, como: "*is a subtype of*", *"[attribute] has value*" (procedure site), entre outros. O conjunto de relações pode ser suficiente para definir o conceito, que pode possuir o status "definido". Quando não há essa suficiência para definição, o conceito possui o status "primitivo". É considerada ainda uma terminologia projetada para permitir o mapeamento de seus conceitos para sistemas de classificação como a CID.

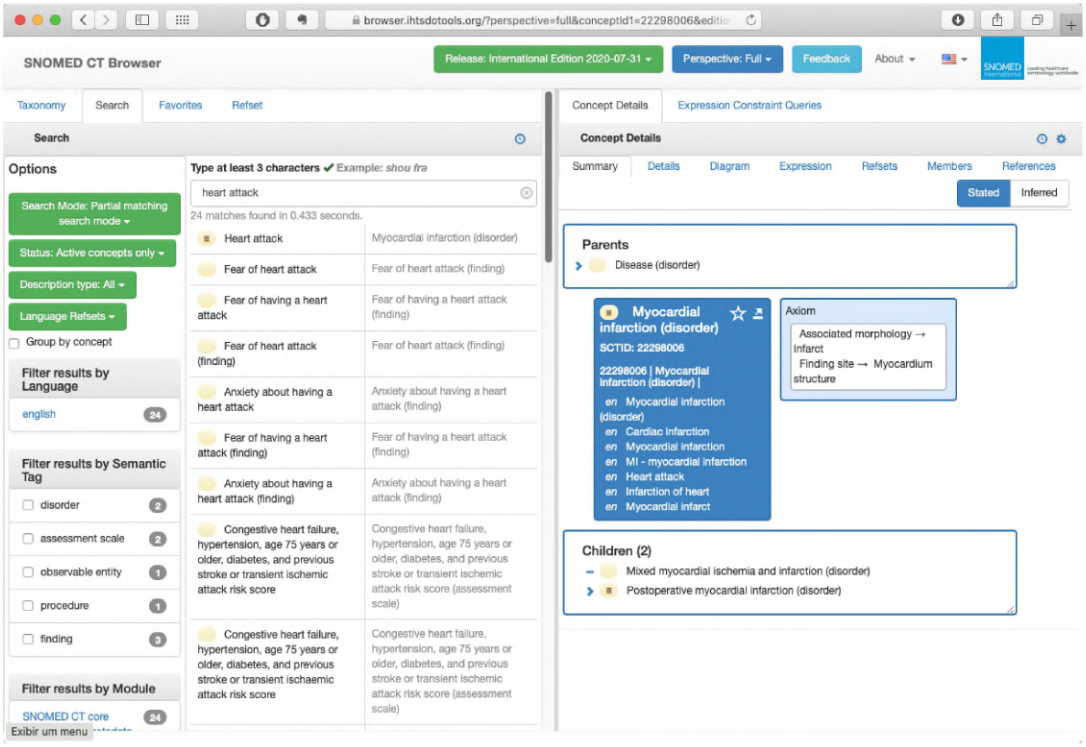
Este conjunto de registros permite ainda uma visão retrospectiva na busca por padrões de ocorrência e/ou tratamentos que requerem acompanhamento adicional.

Deste modo, a abordagem para o manuseio da terminologia SNOMED exige um algoritmo distinto das terminologias MeSH e DeCS. Sua distribuição pela MLDS é feita no formato *Release Format 2* (RF2[20]), codificado em UTF-8[21], no estilo *plain text* (texto puro, sem formatação) e informações em registros (linhas) separadas por tabulação (TSV[22] - *Tab-Separated Values*). Sua transposição para um modelo relacional obedece a mesma modelagem de dados publicada nos arquivos em texto puro. Sua estrutura hierárquica e demais relações conceituais são definidas por axiomas[23] em sintaxe OWL.

---

[20] https://confluence.ihtsdotools.org/display/DOCGLOSS/RF2
[21] https://www.iso.org/standard/63182.html
[22] https://www.iana.org/assignments/media-types/text/tab-separated-values
[23] https://www.w3.org/TR/owl2-syntax/#Axioms

Figura 3 - Interface web da terminologia SNOMED CT



Fonte: https://browser.ihtsdotools.org/?

4 BANCO DE DADOS

Esse capítulo procura apresentar dois paradigmas em estruturas de dados usados neste trabalho.

4.1 Paradigma relacional

O modelo relacional teve início na década 1970, por Edgar Codd, em contraste com os modelos de rede e hierárquicos até então. O modelo apresentado por Codd permitia a descrição dos dados sem a necessidade de estruturas adicionais. A visão relacional possibilitava representações de maior independência dos dados, além de separar a camada de persistência da camada de algoritmos que regem o negócio da aplicação (CODD, 1970).

Sem dúvida o modelo relacional foi muito mais que uma evolução, foi uma revolução que afetou o modo como a indústria de *software* se posicionou com relação aos bancos de dados até hoje. Pode-se considerar que a maior vantagem do modelo relacional sobre seus antecessores, é a representação simples dos dados e a facilidade com que consultas complexas podem ser submetidas.

A fim de manipular o modelo relacional e oferecer à indústria uma linguagem de forma padronizada, a IBM desenvolveu a linguagem estruturada de consultas - SQL[24], denominada inicialmente Sequel, como parte do projeto System R ainda na década de 1970. Estavam à frente deste projeto Donald Chamberlin e Ray Boyce. A linguagem foi projetada a partir do artigo de Codd (1970) "*A Relational Model of Data for Large Shared Data Banks*". A intenção era permitir que pessoas pudessem manipular bancos de dados sem necessariamente conhecer os fundamentos matemáticos (álgebra relacional) ou possuir formação na computação. Mais tarde a linguagem se tornou o padrão para manipulação e consultas em SGBDs relacionais (KELECHAVA, 2018).

Em 1986, a *American National Standards Institute* (ANSI) e a *International Organization for Standardization* (ISO) publicaram um padrão SQL chamado SQL-86. Em 1989 a ANSI publicou uma extensão para a linguagem, chamada SQL-89, de modo que outras versões se seguiram: SQL-92, SQL:1999, SQL:2003, SQL:2006, SQL:2008, SQL:2011 e SQL:2016 (SILBERSCHATZ; KORTH; SUDARSHAN, 2020). A adoção de uma linguagem

[24] Structured Query Language

padrão permitiu aos usuários uma rápida adaptação na necessidade de trabalhar com *softwares* de diferentes fabricantes. Embora alguns fabricantes criassem comandos e possibilidades de extensão da linguagem, a exemplo da PL/SQL da fabricante Oracle, a maioria dos *softwares* adotaram sua forma padrão.

De acordo com Elmasri e Navathe (2011), a fim de possibilitar integridade aos dados, os SGBDs relacionais utilizam o conceito ACID, cujas propriedades são:

- Atomicidade – Uma transação deve ser executada em sua totalidade ou não ser executada.
- Consistência – Uma transação deve ser completamente executada do início ao fim, sem interferência de outras transações.
- Isolamento – Uma transação deve ser executada como de forma isolada, sem interferência de outras transações, mesmo sendo executadas simultaneamente.
- Durabilidade – As alterações aplicadas ao banco de dados por uma transação devem persistir e não podem ser perdidas por qualquer falha.

Esse conjunto de garantias exige configurações complexas de *software*, comprometendo o desempenho de modo geral. Os SGBDs relacionais evoluíram para oferecer processos de validação, verificação e garantias de integridade de dados, controle de concorrência, recuperação de falhas, controle de transações, otimização de *queries*, entre outros (BRITO, 2010). Mas este conjunto de funcionalidades tem um preço computacional e as soluções de SGBDs relacionais em cenários de grande volume de dados e a crescente dificuldade em implementar escalabilidade apontaram a necessidade de criação de novas tecnologias.

4.2 Paradigma NOSQL

As características dos SGBDs relacionais e suas limitações relacionadas às informações não estruturadas, tipo cada vez mais frequentes no ambiente web, suas dificuldades de escalabilidade, entre outras questões, provocou uma procura por alternativas ao modelo relacional. Cientistas da computação ligados à área de BD passaram a desenvolver estratégias de armazenamento desvinculadas de determinadas regras e modelagens que marcaram o modelo relacional. A flexibilização desejada das consistências relacionais poderia permitir um ambiente mais leve, voltado à performance, flexibilização e adaptação dos dados e facilidade de escalabilidade (BRITO, 2010).

No final da década de 1990 o termo NoSQL surgiu como o acrônimo de *Not Only SQL* em banco de dados que não ofereciam interface SQL. O propósito destes *softwares* não é substituir os SGBDs relacionais, mas possibilitar alternativas em aplicações específicas, que demandam a flexibilidade estrutural e demais características deste modelo de dados. Este novo conceito de "consistência eventual" foi aceito inicialmente pelo benefício da performance e escalabilidade, e que era suportado por alguns modelos de negócio, como redes sociais por exemplo (SILBERSCHATZ; KORTH; SUDARSHAN, 2020). Nos anos 2000 iniciava-se uma série de desenvolvimentos a respeito deste paradigma, que de certa forma, retoma o cenário que antecede os SGBDs relacionais, como os bancos de dados hierárquicos. *Softwares* como Bigtable (Google), Apache HBase, Dynamo (Amazon), Cassandra (Facebook), MongoDB, Azure cloud, Sherpa/PNUTS (Yahoo) e Elasticsearch, são alguns exemplos de bancos NoSQL desenvolvidos a partir daquele período.

Segundo Brito (2010) e Toth (2011), os tipos de banco de dados NoSQL podem ser categorizados em:

a) Chave-valor – Coleção de chaves únicas e seus respectivos valores;
b) Orientado a Documentos – Documentos são unidades básicas e não dependem de estrutura pré-definida. Segue o formato JSON[25] para representação. Este é o modelo usado pelo *Elasticsearch*, e explorado neste trabalho.
c) Orientado a Coluna – Constitui uma inversão em comparação dos registros (linhas) no modelo relacional, para os atributos (colunas) neste modelo.
d) Baseado em Grafos – As informações são armazenadas em nós de grafos, onde as arestas são as associações entre os nós, como por exemplo o Neo4J[26].

Uma característica importante relativa ao armazenamento baseado em chave-valor, está na capacidade de lidar com grandes quantidades de dados e consultas, permitindo que estruturas sejam distribuídas em *clusters,* em um grande número de máquinas. Os registros podem ser particionados entre as máquinas no cluster, onde cada máquina pode armazenar um subconjunto dos registros para processamento de *queries* e atualizações. Vários sistemas neste modelo permitem que os dados a serem armazenados sigam uma representação específica, permitindo que o sistema interprete os valores e execute consultas com base no conjunto armazenado, esse modelo é chamado orientado a documentos (SILBERSCHATZ; KORTH; SUDARSHAN, 2020).

[25] JavaScript Object Notation – um formato compacto, para troca de dados simples entre sistemas

[26] https://neo4j.com

4.3 Expansão de *queries*

A Recuperação da Informação enquanto modelo de representação de uma necessidade e a consequente avaliação de um conjunto de respostas, é o fundamento onde os usuários procuram informações, sejam em sistemas especialistas em suas empresas, bibliotecas, ou mesmo na Internet.

A formulação da consulta por usuários não técnicos, geralmente é representada com duas ou três palavras relacionadas ao tópico de interesse. Embora ajustes possam ser feitos para que novas consultas sejam realizadas, muitas vezes os usuários tendem a analisar resultados iniciais a fim de selecionar documentos que atendam sua demanda. Em ambientes como a Internet, onde há trilhões de documentos à disposição, indexados por processos automáticos, consultas formuladas de modo curto, geralmente resulta em uma grande revocação, seguida de baixa precisão.

A necessidade em formular consultas (*queries)* mais específicas é uma realidade, principalmente em grandes coleções e acervos de áreas específicas (WALKER, 2001). A expansão de *queries* é uma técnica de aprimoramento na revocação de resultados, a fim de superar problemas causados por representação conceitual em termos diferentes (VOORHESS, 1994) (XU; CROFT, 1996). Há naturalmente a preocupação de que, com o acréscimo de mais termos, originados das terminologias, sejam tesauros ou ontologias, estes sistemas de recuperação devem aumentar sua revocação, o que pode comprometer a precisão. Mas partindo do pressuposto que os termos relacionados na terminologia possuem boa conexão semântica, estes termos devem acrescentar documentos relevantes no conjunto de resposta.

Qiu e Frei (1993) perceberam os seguintes métodos para expansão de *queries*: a) o uso de coocorrência por similaridade - as similaridades entre os termos são calculadas com base na associação de hipótese, que em seguida são classificados e atribuídos pesos. São criadas estruturas de índices por classes de termos. A consulta é expandida adicionando todos os termos das classes que contêm termos da classe da consulta, b) classificação de documento - os documentos são classificados com um algoritmo de classificação, termos são agrupados em classes e a consulta é expandida substituindo o termo por um conjunto de termos da classe, c) uso de contexto sintático - as relações entre os termos são geradas com base no conhecimento linguístico e nas estatísticas de coocorrência. Este método usa gramática e dicionário para extrair termos de uma lista a fim de expandir a consulta, d) uso de informações relevantes - as informações relevantes são usadas para construir uma estrutura global de informações. A consulta é expandida com essa estrutura de informações.

Já segundo Walker (2001) há dois métodos de expansão: a) *feedback* por relevância - onde o usuário seleciona palavras ou documentos que considere relevante para ampliar a consulta, b) expansão automática - onde o sistema adiciona termos sem a escolha do usuário. Este trabalho deverá utilizar como modelo, o método automático de expansão de *queries*.

A expansão de *queries* em sua forma semiautomática pode ser vista da seguinte forma: um usuário formula sua consulta e submete-a ao sistema de recuperação, um conjunto inicial de respostas lhe é devolvido. O usuário seleciona deste conjunto inicial, os documentos mais adequados à resposta desejada. Estes documentos selecionados são usados pelo sistema a fim de realimentar a consulta e recuperar novos documentos, expandindo o conjunto inicial pela comparação de proximidade entre documentos que inicialmente não respondiam à primeira consulta. Neste exemplo, não há conexão entre a expansão de *queries* e instrumentos terminológicos, como tesauros ou ontologias. Mas outra possibilidade encontra-se no processo de sugestão de outros termos a partir do conjunto de documentos recuperados inicialmente, com apoio de dicionário de sinônimos ou análise de extração em tempo real (XU; CROFT, 1996).

A forma automática de expansão apoiada em instrumentos terminológicos, possui como característica principal, a proximidade conceitual entre os termos, as relações de transitividade hierárquica, e conexões por declaração não hierárquica. A partir de um ou vários termos identificados no texto da consulta inicial, pode-se agregar novos termos a partir de uma terminologia (KOS). Essa consulta, portanto, expandida com novos termos, sinônimos ou conceitualmente próximos, pode resultar em resultados mais amplos do que a consulta original.

A questão de linguagem é a situação mais crítica na efetividade do processo de RI, é o problema de incompatibilidade entre a representação pelos termos do usuário em comparação com a representação dos indexadores e/ou autores. Acrescenta-se a essa questão, problemas linguísticos como, sinônimos, polissemia, variações entre singular e plural, entre outros desafios. Uma série de técnicas foram desenvolvidas para lidar com estas situações, pode-se citar: refinamento de *queries*, *feedback* por relevância, desambiguação de sentido de palavras e agrupamento de resultados de pesquisa (clusterização) (CARPINETO; ROMANO, 2012). Portanto, uma das técnicas mais naturais para a expandir a consulta (*query*) original, é incluir termos que traduzam o mesmo sentido da necessidade do usuário em diferentes representações (grafias), a fim de melhorar os resultados pela comparação com a representação dos documentos pesquisados.

O uso de relações estatísticas para expansão de *queries*, baseado no modelo vetorial é interessante, pois as relações são geradas a partir dos documentos já existentes, evitando o custo de criação e manutenção de terminologias. Mas estes métodos tiveram pouco sucesso na

melhoria da RI, quando usados em dados de pouca relevância no contexto do acervo (VOORHESS, 1994).

Salton (1983) e Lesk (1996) pesquisaram que a expansão por sinônimos melhorou o desempenho na RI, mas a expansão por termos mais restritos ou selecionados por estrutura de sinônimos hierárquicos geraram dados inconsistentes. Esta situação pode ser efeito de uma expansão baseada em termos pobres conceitualmente, dentro do universo da área de conhecimento específica. Outra experiência negativa pode ser lida em Voorhess (1994). Seu trabalho usou como estratégia a expansão de consulta com sinônimos e seus descendentes em uma relação "é um" (*is-a)*. Inicialmente sua avaliação do experimento apontou que não houve melhora significativa com o desempenho da consulta expandida, na avaliação do autor o diferencial foi considerado muito pequeno para a maioria das consultas realizadas. Também foi identificado que consultas mais longas (com mais palavras) são menos eficientes que consultas mais curtas. Consultas mais longas devem ser aprimoradas, a exemplo de técnicas de *feedback* por relevância. Percebe-se ainda que a relações mais úteis para a expansão de *queries* está diretamente associada à capacidade de associação conceitual à área específica do objeto procurado. Em contraste, o experimento de Qiu e Frei (1993) relata um experimento de sucesso, relacionado a expansão automática de *queries*. Naquele trabalho, a criação do tesauro automático em sua extensão e qualidade terminológica, foi o ponto diferencial para que o processo de expansão tivesse expressiva recuperação dos documentos.

Especificamente sobre a expansão de *queries* e seu uso com apoio de SOCs, pode-se ressaltar pesquisas como Qiu e Frei (1993), Walker (2001), Müller, Kenny e Sternberg (2004), Shiri e Revie (2006), Hollink, Malaisé e Schreiber (2010) com enfoque nos tesauros, Voorhees (1994) e as relações lexicais semânticas, Aronson e Rindflesch (1997) com UMLS Metathesaurus, Efthimiadis (2000) e avaliação por feedback de usuários, Abdou e Savoy (2008) com Medline e MeSH, Stocker *et al*. (2018) no domínio do genoma, Sánchez e Azpilicueta (2011), Zivaljevic, Atalag e Warren (2020) no uso de ontologias, Carpineto e Romano (2012), Azad e Deepak (2019) em uma ampla discussão sobre o tema. Percebe-se, portanto, o contínuo interesse do tema nas publicações científicas, fator indicativo da necessidade de continuidade de discussão e evolução sobre essa área de pesquisa.

Parte - II

5 METODOLOGIA

5.1 Metodologia de pesquisa científica

Os capítulos anteriores permitiram uma visão geral dos temas associados à pesquisa, bem como as principais referências da literatura sobre a área de conhecimento envolvida. Após a introdução, o capítulo 2 apresentou a Recuperação da Informação e suas relações com os sistemas de organização do conhecimento, especificamente, tesauros e ontologias. Descreveram-se ainda métricas que são os verdadeiros desafios da RI, a saber, a revocação e precisão, finalizando-se com a expansão de *consultas*, técnica já adotada em sistemas de RI desde há muito e que ainda se mostra eficiente para ampliar as possibilidades de recuperação semântica. No capítulo 3, enfatizaram-se aspectos dos artefatos terminológicos usados no desenvolvimento do experimento com o *software* nesta pesquisa, bem como particularidades no domínio de conhecimento da medicina.

Na questão do domínio de conhecimento, o trabalho se desenvolve como uma interseção entre as áreas de Ciências Sociais Aplicadas e de Ciências Exatas e da Terra. Do ponto de vista da metodologia de pesquisa, o presente trabalho pode ser classificado da seguinte forma (GIL, 2010):

- Quanto à sua finalidade: pesquisa aplicada;
- Quanto aos seus objetivos: pesquisa exploratória, pois envolve o desenvolvimento de algoritmos para aplicação, bem como a exploração e uso de artefatos terminológicos na área médica;
- Quanto à natureza dos dados: pesquisa qualitativa;
- Quanto a modalidade: estudo de caso, por realizar uma aplicação de situação de experimento prático na condução da análise de dados.

O presente capítulo dedica-se a detalhar os passos metodológicos na condução do experimento, bem como seus desdobramentos. A Seção 5.2 aponta os princípios que nortearam a pesquisa e o *software* desenvolvido, bem como os passos metodológicos: como foi definida a coleta de dados, os problemas e desvios encontrados, além do desenvolvimento do experimento e suas limitações.

5.2 Metodologia da pesquisa

A presente seção descreve o contexto da pesquisa e os passos metodológicos, os quais são apresentados em detalhe ao longo das subseções correspondentes.

5.2.1 Contexto da pesquisa

A constatação de que a Recuperação da Informação de documentos se estabelece com a comparação de símbolos nos textos motivou a investigação. Na verdade, a representação dos símbolos textuais em computadores, sua identificação, o estabelecimento da correspondência entre a requisição do usuário e a representação do documento é um dos aspectos mais relevantes da RI. Diversas técnicas foram desenvolvidas tendo como base esse princípio de comparação, como apresenta-se ao longo deste capítulo.

Visto que os conceitos podem ser representados por símbolos diferentes, além dos conceitos relacionados por sinonímia, alcançam-se as variações semânticas. No contexto dos computadores, o estudo de aspectos sintáticos e semânticos se revela uma contribuição importante fornecida pelos artefatos terminológicos – no presente trabalho, terminologias médicas especializadas – uma vez que permitem a expansão da consulta no âmbito do processo de RI.

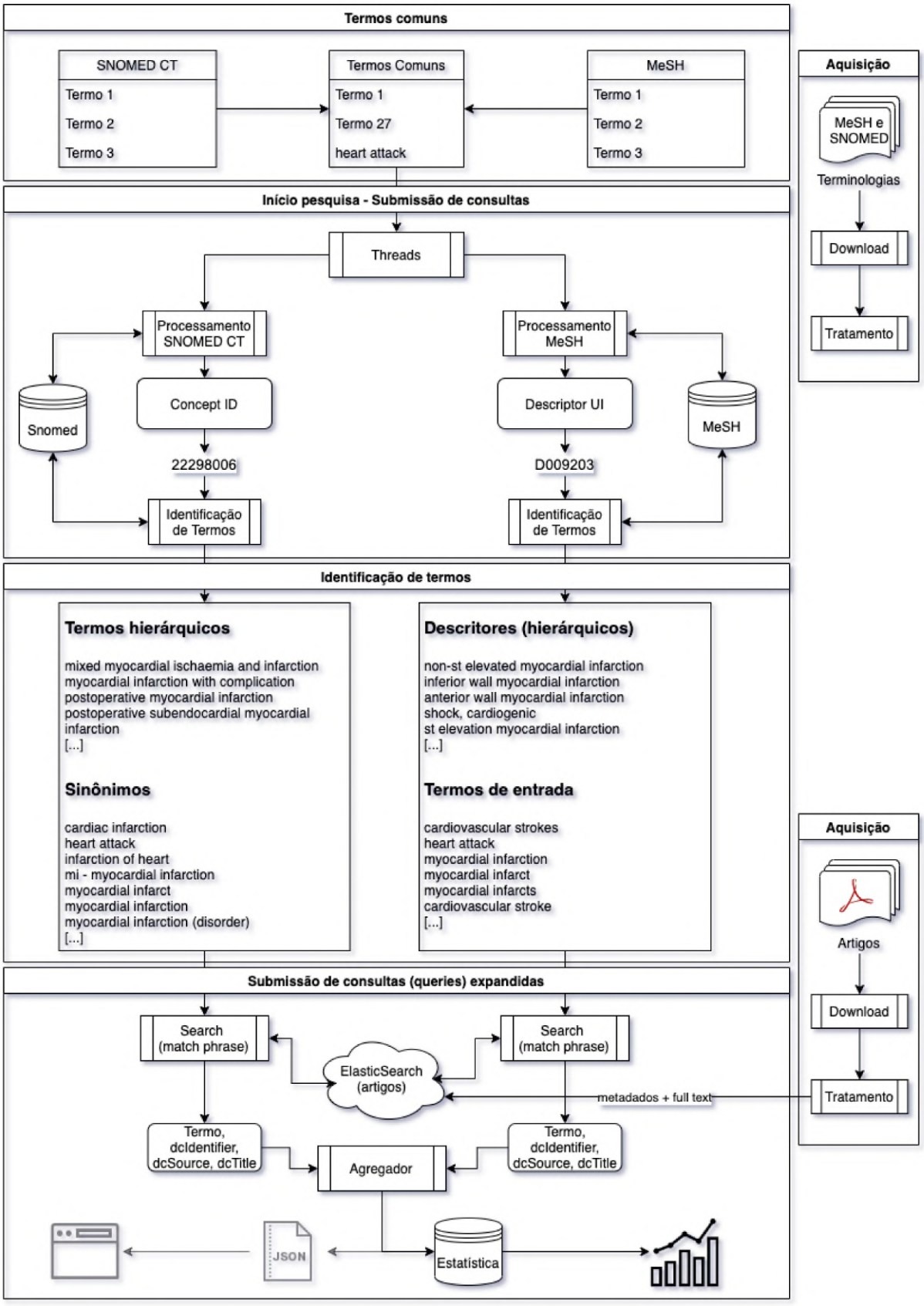
No âmbito desses princípios, podem-se observar limitações dos processos de RI principalmente em ambientes onde o vocabulário é técnico e restrito ao grupo especializado. É comum nesse tipo de ambiente a ocorrência de siglas, abreviações e termos técnicos diversos que convergem para significado de um conceito. Para isso, os sistemas de representação do conhecimento como taxonomias, tesauros e ontologia contribuem para melhorias na RI. As possibilidades de relação terminológicas encontradas nesses artefatos, com seus diferentes níveis de expressividade, permitem que termos diferentes se relacionem para representar o mesmo conceito.

5.2.2 Passos metodológicos

A presente seção é organizada para descrever os cinco passos iniciais e os passos do experimento em si, apresentados na seguinte ordem: i) *passos inicias*; e ii) *passos do experimento*. Na sequência, cada um dos passos inicias e do próprio experimento são detalhados para melhor entendimento dos processos realizados.

A figura 4 procura mostrar as principais etapas de forma gráfica. Estas etapas serão exploradas nos próximos tópicos com maior detalhe descritivo.

Figura 4 - Processo do experimento de software



Fonte: Dados da pesquisa

5.2.2.1 Os passos iniciais

Os passos iniciais do experimento a serem detalhados aqui são:

A. Aquisição das terminologias;
B. Estruturação terminológica;
C. Aquisição de artigos;
D. Pré-processamento;
E. Estruturação dos artigos;

Esses cinco passos com seus algoritmos periféricos compuseram todo o experimento e foram fundamentais para que o algoritmo principal pudesse processar as consultas e gerar os resultados.

Como ferramentas de apoio, destacam-se ainda:

i. Github, para armazenamento, backup e versionamento dos algoritmos;
ii. Docker, para permitir a instalação do banco de dados para os artigos no modelo *container,* e facilitar sua instalação e manutenção;
iii. *Elasticsearch*, banco de dados NoSQL para armazenamento do texto e metadados dos artigos, em formato texto completo;
iv. Kibana, interface para acesso em alto nível da suíte de soluções *Elasticsearch*, a fim de executar *scripts*, consultas, e permitir a manutenção do banco de dados;
v. SQLite, banco de dados SQL para armazenar as terminologias modeladas no esquema relacional e os resultados obtidos pelas submissões das consultas;
vi. Dbeaver, cliente SQL para administração de diversos banco de dados e usado para gerenciar os arquivos SQLite;
vii. Flask, micro servidor web em Python usado no projeto de interface com o usuário;
viii. Trello, para apoio na gestão de tarefas, definição de prioridades e controle de processos;
ix. Visual Studio Code, para edição de código fonte no desenvolvimento de *software*.

A. Aquisição das terminologias

O procedimento para acesso e uso dos artefatos foi realizado no meio digital, a partir do preenchimento de cadastro no site de cada instituição.

Surgiram, portanto, diversas iniciativas em vocabulários médicos já na década de 1950. Cita-se abaixo, alguns vocabulários relevantes no contexto médico:

- *International Classification of Diseases* (ICD[27]): iniciativa de 1893, atualmente é um sistema de classificação para permitir o processamento de informações sobre doenças.
- *Medical Subject Headings* (MeSH): detalhado na seção 3.1.2.
- *Systematized Nomenclature of Medicine* (SNOMED): detalhado na seção 3.1.3.
- *International Classification of Primary Care* (ICPC[28]): classificação internacional direcionada para informações a respeito de atendimento clínico na atenção primária.

Após a análise para definição dos artefatos terminológicos mais adequados para uso neste projeto, foram escolhidos dois vocabulários: *Systematized Nomenclature of Medicine - Clinical Terms* (SNOMED CT), *Medical Subject Headings* (MeSH). As terminologias foram selecionadas a partir dos seguintes critérios:

- Credibilidade;
- Abrangência semântica em diversas áreas do conhecimento médico;
- Disponibilidade;
- Política de atualização.

Na MeSH, o processo de solicitação de acesso conta com uma análise para autorização após o preenchimento do formulário eletrônico. A *U.S. National Library of Medicine* (NLM) é a instituição responsável pela terminologia e sua política inclui atualizações semanais em diversos formatos de distribuição. A instituição[29] mantém na web diversos *links* para páginas de orientações gerais, documentação e *download* da terminologia. À época do contato inicial, a NLM concedeu acesso à terminologia via FTP[30], o que permite encontrar versões anteriores da terminologia, arquivos de atualização com novos termos e publicações da terminologia em vários idiomas. A aquisição dos artefatos terminológicos, dessa forma, se deu em umas poucas etapas:

a) Solicitação de acesso por contato eletrônico;
b) Aprovação e divulgação de acesso via FTP;
c) Download da terminologia em formato XML,

A terminologia usada está no formato de publicação XML, no ano de 2020.

Para o SNOMED CT há uma licença denominada *Member Licensing and Distribution Service* (MLDS), que pode ser requisitada por formulário eletrônico a ser analisado pela instituição. Após a aprovação, acessou-se um painel de controle (*dashboard*), em uma área

[27] Disponível em: https://www.who.int/classifications/icd/en/
[28] Disponível em: https://www.who.int/classifications/icd/adaptations/icpc2/en/
[29] Disponível na internet em: https://www.nlm.nih.gov/mesh/meshhome.html
[30] File Transfer Protocol

de gerenciamento via web browser, onde o cadastro do pesquisador foi realizado. Como passos para aquisição do SNOMED, pode-se elencar:

a) Solicitação de acesso por formulário eletrônico;
b) Aprovação e divulgação de acesso via *dashboard*;
c) Download da terminologia em formato RF2;

Há no mínimo duas atualizações da terminologia por ano, que podem ser acessadas diretamente pelo painel de controle, bem como diversos documentos. No presente projeto, o código de filiação do pesquisador foi 301455, tipo “acadêmico” e subtipo “pesquisa”.

Uma vez de posse das duas terminologias clínicas SNOMED CT e MeSH, e suas respectivas autorizações de uso, os procedimentos de padronização para interação com as mesmas iniciou-se. Na seção 5.2.2.2 serão esclarecidos os procedimentos de utilização das terminologias.

B. Estruturação terminológica

As terminologias são publicadas originalmente em formatos que nem sempre são adequados para manipulação, como é o caso daquela planejada para o experimento do presente trabalho.

Dentre os formatos disponíveis da terminologia MeSH, o arquivo escolhido para o processamento está codificado em XML. Após a análise dos dados pertinentes ao projeto, foram identificados os seguintes campos e mapeamentos originais:

Quadro 2 - Principais campos e mapeamentos originais MeSH

Formato original	Informação	Descrição
<DescriptorRecordSet LanguageCode = "eng">	eng	Idioma da publicação
<DescriptorUI>D000001</DescriptorUI>	D000001	Identificador do descritor, campo chave.
<DescriptorName> <String>Calcimycin</String> </DescriptorName>	Calcimycin	Nome do descritor principal de determinado conceito
<TreeNumberList> <TreeNumber> D03.633.100.221.173 </TreeNumber> </TreeNumberList>	D03.633.100.221.173	Código que define a posição hierárquica do descritor

<TermList> <Term ConceptPreferredTermYN="N" IsPermutedTermYN="N" LexicalTag="NON" RecordPreferredTermYN="N"> <TermUI>T000003</TermUI> <String>Antibiotic A23187</String>	T000003 e Antibiotic A23187	Códigos e termos de entrada relacionados ao descritor principal. Um descritor pode ter uma lista de outros termos, elencados na tag <TermList>

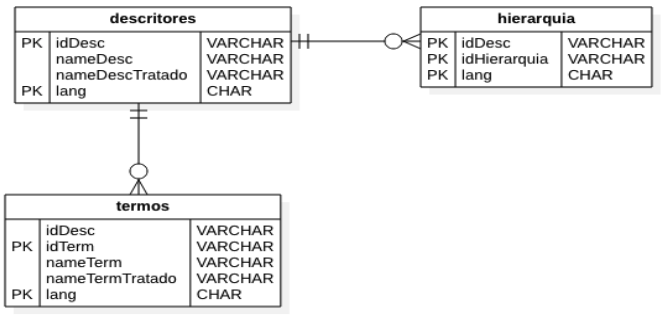
Fonte: Dados da pesquisa

O conjunto de informações (quadro 2) foi mapeado para o modelo relacional da seguinte forma:

- Um descritor pode ter vários termos de entrada;
- Um descritor pode ter vários códigos de hierarquia;

A partir desse modelo, foram criadas três tabelas com os respectivos relacionamentos (Figura 5). Usando um algoritmo escrito em linguagem Python, foi realizada a carga de dados no banco de dados SQLite com as informações extraídas do arquivo XML. O dicionário de dados do modelo também foi confeccionado (quadro 3).

Figura 5 - Modelo lógico para a terminologia MeSH



Fonte: Dados da pesquisa

Quadro 3 - Dicionário de dados do modelo MeSH

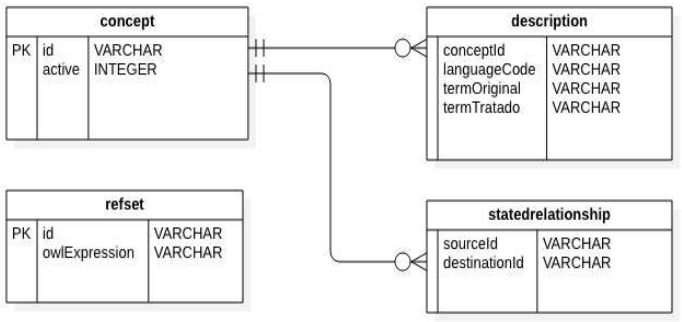
Campo	Descrição
idDesc	Identificador do texto descritor
nameDesc	Nome do descritor
nameDescTratado	Descritor tratado para retirada de caracteres especiais
lang	Idioma
idHierarquia	Identificador do código hierárquico
idTerm	Identificador do termo de entrada
nameTerm	Nome do termo

nameTermTratado	Nome do termo tratado para retirada de caracteres especiais

Fonte: Dados da pesquisa

Raciocínio e procedimentos similares foram usados para a terminologia SNOMED CT. O arquivo publicado pela MLDS se encontra codificado em texto no formato TSV, e a modelagem obedeceu ao padrão já estabelecido pelo arquivo original. Cada arquivo texto deu origem a uma tabela relacional mais seus respectivos campos. Todos os campos foram considerados com sua nomenclatura original. Foram importados quatro dos seis arquivos disponíveis gerando o modelo relacional (Figura 6), o qual é explicado por um dicionário de dados (quadro 4). Dessa forma, foi criado outro banco de dados, específico para a SNOMED, também usando como plataforma de implementação o SQLite.

Figura 6 - Modelo lógico para a terminologia SNOMED CT



Fonte: Dados da pesquisa

Quadro 4 - Dicionário de dados do modelo SNOMED CT

Campo	Descrição
Tabela concept	
id	Identificador do conceito
active	Código para indicar registro ativo ou inativo
Tabela description	
conceptId	Identificador do conceito
languageCode	Idioma
termOriginal	Descrição do termo original
termTratado	Descrição do termo tratado
Tabela statedrelationship	
sourceId	Identificador do conceito (origem)

destinationId	Identificador do conceito (destino)
Tabela refset	
id	Identificador do axioma
owlExpression	Axiomas com definições de hierarquia e relacionamentos

Fonte: Dados da pesquisa

Finalmente, cabe citar que em ambos bancos de dados foram criados índices nos campos críticos de consultas internas, a fim de melhorar a performance de *consultas* mais complexas. Assim, ambos os bancos de dados foram otimizados da seguinte forma: i) forma criados somente os campos necessários; b) foram criados índices nos principais campos. Tais iniciativas planejavam proporcionar consultas rápidas, além de reduzir o tamanho do arquivo em disco contribuindo para o desempenho do *software* no experimento.

C. Aquisição dos artigos

Para realizar o experimento de RI, considerando as terminologias clínicas e a linguagem natural do usuário requisitante, foi feito um levantamento para seleção e aquisição de artigos científicos. Era necessário identificar um conjunto de documentos em uma mesma área temática que pudesse responder ao mesmo domínio de conhecimento das terminologias clínicas em questão. Inicialmente foram definidas duas áreas temáticas: Obstetrícia ou Geriatria, as quais são ligadas ao grupo de pesquisa[31].

A pesquisa por acervos de artigos foi feita utilizando buscador na Internet, usando como termo de busca as seguintes frases: i) "*medical articles*"; ii) "*biomedical articles*".

Entre os acervos encontrados, a escolha dos artigos foi decidida pelos seguintes fatores:

- Disponibilidade temática;
- Publicação dos artigos em formato completo (full papers);
- Acervo de livre acesso (open access);
- Instituição com credibilidade no meio acadêmico;
- Quantidade razoável de artigos disponíveis.

A quadro 5 mostra o comparativo do acervo encontrado e as principais características:

[31] Desejava-se que a temática tivesse conexão com o grupo de pesquisas ReCOL - http://recol.eci.ufmg.br

Quadro 5 - Comparativos de acervos de artigos médicos

	The New England Journal of Medicine	*BMC*	*ScienceDirect*	*PubMed Central (PMC)*	*thebmj*	*The Medical Journal of Australia*
Disponibilidade temática	Sim	Por categoria	Sim	Sim	Sim	Sim
Artigos completos	Sim	Sim	Sim	Sim	Sim	Sim
Livre acesso	Alguns	Sim	Não	Sim	Não	Não
Credibilidade	Sim	Sim	Sim	Sim	Sim	Sim
Quantidade > 300	Não	Sim	Sim	Sim	Sim	Sim
Referência	https://www.nejm.org/medical-articles/research	https://bmcmedicine.biomedcentral.com/articles	https://www.sciencedirect.com	https://www.ncbi.nlm.nih.gov/pmc/	https://www.bmj.com/	https://www.mja.com.au/

Fonte: Dados da pesquisa.

Diante desse conjunto de critérios, foi identificado o acervo da BMC Medicine[32] como o mais adequado, o qual foi seguido pela PubMed Central. Contando com um grande volume de artigos científicos da área médica em formato completo ("*texto completo*"), o acervo da BMC organiza seus artigos por grandes grupos da área médica. Trata-se de instituição filiada a *Springer Nature*, outra instituição de grande reconhecimento científico. Dentre os temas selecionados para esse trabalho, "Geriatria" estava disponível.

Para baixar os artigos foi desenvolvido um algoritmo para "visitar as páginas" do site do acervo, identificar os endereços web dos artigos (*url*) e realizar seu download sem, contudo, impactar negativamente a performance do servidor de origem. Esse é o processo denominado *web scraping*, que pode ser feito de diversas formas, independente da linguagem de programação. Os primeiros algoritmos para *scraping* foram desenvolvidos para mecanismos de busca (HAJBA, 2018), mas com a demanda para responder às milhares de pesquisas, esses programas evoluíram para indexar a web com o reconhecimento de urls e, nesse sentido, são denominados *crawlers*.

O algoritmo para esse procedimento foi desenvolvido na linguagem *Python* e contou principalmente com o auxílio dos recursos das bibliotecas *request* e *beautifulsoup*. A primeira biblioteca possui uma coleção de métodos que atendem ao HTTP e recursos para serializar dados entre o servidor e o cliente. Nesse trabalho, a biblioteca *request* foi utilizada

[32] https://bmcmedicine.biomedcentral.com/articles

para o processo de requisição ao endereço web, retornando o conteúdo da página. A segunda biblioteca, *beautifulsoup,* foi utilizada no processo de análise dos dados (*parse*) e permitiu a identificação dos endereços de artigos em formato pdf para que, novamente, por meio da biblioteca *request* fosse realizada a requisição para *download* de cada artigo.

Deste modo, realizou-se o *download* dos artigos, totalizando 2.481 arquivos em formato pdf. Os resultados desse processo estão descritos em maiores detalhes na seção 6.1.1.3.

D. Pré-processamento

Após a coleta de dados, foi necessária uma completa transformação do conteúdo dos artigos originais disponíveis no formato pdf. A importância dessa preparação – o pré-processamento – é fundamental para a manipulação dos dados pelos algoritmos. Tendo em vista que não se tinha acesso ao material original, à saber, o texto sem formatação, mas sim ao formato resultante da renderização em pdf, era necessário realizar o processo inverso. O resultado foi um formato de texto passível de manipulação pelo banco de dados e linguagem de programação.

Os critérios usados para a avaliação e escolha da biblioteca mais adequada foram: i) quantidade de caracteres; ii) quantidade de *tokens*, iii) tempo de extração, iv) qualidade[33] de extração, v) ordem lógica textual,[34] vi) formato adequado para RI. Dentre os critérios ressalta-se a importância dos três últimos itens[35]. O procedimento fez uso de 6 artigos médicos em formato pdf, escolhidos de forma controlada, para obter arquivos de diferentes datas de edição e conversão. Esse critério, contudo, revelou a impossibilidade de uso de uma das bibliotecas que, a partir da leitura dos arquivos de 2016 e 2019, não retornou qualidade de extração adequada, unificando palavras e produzindo um texto sem sentido.

Tabela 1 - Amostras de pré-processamento nos artigos

Arquivo 1471-2318-2-3.pdf - artigo de 2002 - 6 paginas com 2 figuras e 5 tabelas				
	pymupdf	pdftotext	Tika-Python	PyPDF2
Caracteres	24,479	39,762	30,664	24,469

[33] Remete a integridade das palavras e dos caracteres originais do texto, como símbolos especiais e acentuação.
[34] A ordem lógica textual diz respeito ao fluxo da mensagem: considera-se como formato adequado para RI um texto que, desconsiderando quebras de linhas e submetido a pré-tratamento, preserve a sequência de palavras e continue fiel à informação contida no arquivo original.
[35] O algoritmo usado na comparação está disponível em: https://github.com/erfelipe/PDFtextExtraction

Tokens	4,268	4,268	5,255	4,116
Tempo extração	0.119519146	0.082923195	17.58037679	0.543136894
Qualidade extração*[36]	Ótima	Ótima	Ótima	Boa
Ordem lógica textual	Sim	Não	Sim	Sim
Formato adequado para RI	Adequado	Inadequada	Adequado	Adequado
Arquivo 1471-2318-7-23.pdf - artigo de 2007 - 12 paginas com 1 figura e 3 tabelas				
	pymupdf	pdftotext	Tika-Python	PyPDF2
Caracteres	55,197	73,010	65,511	55,013
Tokens	8,684	8,686	10,266	8,566
Tempo extração	0.094203851	0.131608824	0.582838876	0.556111888
Qualidade extração*	Ótima	Ótima	Ótima	Boa
Ordem lógica textual	Sim	Não	Sim	Sim
Formato adequado para RI	Adequado	Inadequada	Adequado	Adequado
Arquivo 1471-2318-9-12.pdf - artigo de 2009 - 9 paginas com 2 figuras e 3 tabelas				
	pymupdf	pdftotext	Tika-Python	PyPDF2
Caracteres	44,204	56,925	52,665	44,555
Tokens	7,153	7,155	8,495	7,139
Tempo extração	0.072447393	0.116884809	0.501214137	0.418291169
Qualidade extração*	Ótima	Ótima	Ótima	Boa
Ordem lógica textual	Sim	Não	Sim	Sim
Formato adequado para RI	Adequado	Inadequado	Adequado	Adequado
Arquivo s12877-016-0361-8.pdf - artigo de 2016 - 14 paginas com 1 figura				

[36] * Inaceitável - Regular - Boa - Ótima

	pymupdf	pdftotext	Tika-Python	PyPDF2
Caracteres	76,019	88,668	80,012	65,546
Tokens	12,620	12,620	13,175	3,297
Tempo extração	0.090397175	0.158302334	0.548722008	0.584170154
Qualidade extração*	Ótima	Ótima	Ótima	Inaceitável
Ordem lógica textual	Sim	Não	Sim	Sim
Formato adequado para RI	Adequado	Inadequado	Adequado	Inadequado
Arquivo s12877-019-1283-z.pdf - artigo de 2019 - 8 paginas com 1 figura e 4 tabelas				
	pymupdf	pdftotext	Tika-Python	PyPDF2
Caracteres	35,141	49,480	36,228	30,780
Tokens	6,040	6,046	6,143	2,836
Tempo extração	0.077984333	0.135750933	0.520617036	1.234345989
Qualidade extração*	Ótima	Ótima	Ótima	Inaceitável
Ordem lógica textual	Sim	Não	Sim	Sim
Formato adequado para RI	Adequado	Inadequado	Adequado	Inadequado
Arquivo s12877-019-1372-z.pdf - artigo de 2019 - 12 paginas com 1 figura				
	pymupdf	pdftotext	Tika-Python	PyPDF2
Caracteres	70,278	81,671	73,986	61,155
Tokens	11,807	11,809	12,319	3,569
Tempo extração	0.105335754	0.213534637	0.33198551	0.870980965
Qualidade extração*	Ótima	Ótima	Ótima	Inaceitável
Ordem lógica textual	Sim	Não	Sim	Sim
Formato adequado para RI	Adequado	Inadequado	Adequado	Inadequado

Fonte: Dados da pesquisa.

A tabela 1 mostra os resultados dos testes realizados. Das bibliotecas avaliadas, cabe mencionar algumas características:

- *PyMuPDF*: conversão adequada mesmo considerando a existência de tabelas no arquivo pdf; não considera espaços de linhas em branco, o que ajuda no pré-tratamento; tempo de conversão rápido em comparação com outras bibliotecas.
- *Pdftotext*: conversão adequada, a mais fiel dentre as avaliadas, porém extrai o texto em duas colunas como na diagramação original; essa característica gera erro pela junção de frases diferentes.
- *Tika-Python*: conversão adequada, com destaque para o reconhecimento de URLs; contudo, considera muitos espaços de linha em branco, o que acrescenta mais um processo no pré-tratamento aumentando o tempo de processamento; código nativo não é codificado em Python.
- *PyPDF2*: resultou em quebras de linha, o que não ocorreu em outros conversores; em três arquivos do teste realizado, particularmente, os datados em 2016 e 2019, a extração foi inaceitável pela total ausência de espaços entre as palavras.

Portanto, a partir dessa parte pré-experimento, a escolha recai sobre as bibliotecas *PyMuPDF* ou *Tika-Python*. Após a definição da biblioteca de extração mais adequada, os problemas do texto resultante dos artigos em formato pdf eram:

- Cabeçalho e rodapé incluídos como contaminação do corpo textual;
- Ausência de identificação do título como metadado;
- URLs sem formatação adequada;
- Caracteres sem sentido decorrentes da extração de tabelas;
- Caracteres sem sentido decorrentes da dificuldade de identificação do texto.

Outros resultados decorrentes da extração não se apresentam como erro, mas foram tratadas a fim de melhorar o conjunto textual para gravar no banco de dados e para possibilitar RI. Exemplos são a quebra de linha e a padronização em caracteres para minúsculo.

Após a conversão do artigo original em pdf para texto simples, o processo de preparação do texto final a ser gravado no banco de dados foi realizado em etapas, onde diversas técnicas foram aplicadas até que o formato final estivesse compatível com a grafia das terminologias clínicas. Assim, é possível a comparação terminológica e a consequente identificação dos termos com sucesso. A sequência de procedimentos de intervenções no texto – o *pipeline* – é detalhada a seguir em etapas:

- Exclusão de caracteres especiais;
- Tratamento de quebras de linha;

- Tratamento de caixa alta e baixa;
- Remoção de termos desnecessários (*stopwords*);
- Tratamento de caracteres numéricos;
- Codificação;
- Tokenização.

A exclusão de caracteres especiais foi necessária porque ambas terminologias clínicas, estão em inglês. Embora os artigos usados também sejam escritos nesse mesmo idioma, vários caracteres não são necessários na constituição de cada registro do banco de dados, a exemplo de caracteres especiais e símbolos de pontuação. Esse foi o primeiro passo na preparação textual considerando que ambas terminologias não adotam tais caracteres em seus termos. A exclusão permite que seja armazenado um texto limpo, facilitando a pesquisa. Por exemplo, o apóstrofe exerce grande importância na abreviação de termos em inglês, mas no contexto da presente pesquisa não afeta a comparação terminológica e deve ser desconsiderado na preparação do texto final.

Após o processo de conversão, o texto obedece a formatação original de quebras de linha, incluindo a hifenização das palavras. A quebra de linha faz sentido para que o usuário leia o texto de forma agradável em diagramação adequada e é por vezes padronizada com modelos para o ambiente científico. Porém, como referência de consulta, a quebra de linha não é necessária e adiciona complexidade na comparação terminológica. Retirar a quebra de linha ajudou a desenvolver um algoritmo preciso na transformação de palavras hifenizadas para sua sintaxe habitual.

Os caracteres apresentados em maiúsculas ou minúsculas são relevantes em contextos onde a análise textual precisa diferenciá-las para identificar expressões de emoção ou tópicos para separação em agrupamentos temáticos. Nesta pesquisa, é importante que a técnica utilizada para comparar termos dos instrumentos com a base de dados padronize os caracteres para a correta comparação por meio de *case-folding*[37]. De forma que todos os caracteres fiquem em minúsculas, esta técnica é aplicada no corpo do artigo a ser armazenado na base de dados e nos termos de ambas terminologias.

A remoção de *stopwords* é uma técnica muito divulgada que recebeu um certo status de certa obrigatoriedade no pré-processamento textual. Há muitas listas disponíveis na Internet como referência de *stopwords*, e em sua ampla maioria são compostas por artigos, conectivos e termos de pouca expressão semântica. Se considerado o modelo "*bag of words"*, muito usado

[37] https://nlp.stanford.edu/IR-book/html/htmledition/capitalizationcase-folding-1.html

para classificação e Processamento de Linguagem Natural (PLN), a remoção de *stopwords* é importante. No presente trabalho, entretanto, optou-se por não usar a técnica em sua totalidade: artigos "a", "e" e demais conectores, são relevantes porque participam das terminologias clínicas, mas termos como abreviações "fig.", "image", "table", podem ser descartados sem prejuízo para a comparação entre as terminologias e o documento.

Os caracteres numéricos têm importância em pesquisas *em texto completo*, e em contextos onde sistemas de métricas, endereços e medições são relevantes. Nesse trabalho, após uma análise preliminar, identificaram termos nas terminologias que fazem o uso de caracteres numéricos. São, porém, termos muito específicos que por vezes fazem referência a quantidade ou código de medicação. Dessa forma, optou-se por não excluir os caracteres numéricos dos textos a fim de preservar sua sintaxe para comparação com as terminologias.

Existem diversas iniciativas para o mapeamento de caracteres e símbolos como *American Standard Code for Information Interchange* (ASCII), *Unicode Transformation Format* (UTF-8), *American National Standards Institute* (ANSI), Windows-1252, e ISO-8859-1) são padrões que viabilizam a correta exibição de fontes (letras e símbolos) para textos eletrônicos. A maneira como o texto é codificado – denominada *charset* – é importante para proporcionar uma versão final do texto adequada, compreensível e compatível com os mesmos caracteres ou símbolos codificados pelas terminologias clínicas. O padrão de codificação define a forma como o texto resultante da conversão deve ser representado e permite a leitura e o entendimento humano. Considerando o idioma do projeto, onde ambos os artefatos foram codificados em inglês e que os artigos também utilizam o mesmo idioma, não houve necessidade de conversão da codificação.

A tokenização é uma técnica de análise textual que permite a quebra de texto composto por muitas palavras (ou *strings*), em palavras e símbolos (como pontuação, etc.) separados em *substrings* resultando em uma estrutura de dados em forma de lista (*array*) permitindo uma análise individualizada. Nesse trabalho esta técnica foi utilizada como apoio para a escolha da biblioteca de extração pdf, a fim de facilitar a análise comparativa de cada biblioteca.

E. Estruturação de artigos

A escolha do banco de dados para armazenamento dos artigos científicos foi baseada em uma análise com os seguintes aspectos:

- Armazenamento de grande volume textual por item referenciado

- Capacidade de consultas em texto completo (FTS[38]) em um ou vários campos
- Mecanismo de ordenação em *ranking* (desejável)
- Conectividade com a linguagem utilizada no projeto (Python)

Outras características como velocidade de resposta, atomicidade de operações de escrita e ferramentas de automação de backup foram consideradas desejáveis, mas não obrigatórias pois a pesquisa faz uso de um ambiente controlado e sem necessidade de escalabilidade. Também participou do processo de análise, o tipo de estrutura de dados a ser utilizada: se relacional (SQL) ou não relacional (NoSQL). Em particular, esse quesito revelou significativa importância no caso de consultas para texto completo. Três bancos de dados foram selecionados para um comparativo (quadro 6):

Quadro 6 - Comparativo de Banco de Dados

	PostgreSQL	MongoDB	Elasticsearch
Paradigma relacional	Sim	Não	Não
Paradigma orientado a documentos	Não	Sim	Sim
TudoGrande volume textual	Sim	Sim	Sim
Consultas FTS (*texto completo search*)	tsvector	Atlas Search	Nativo
Mecanismo ordenação em ranking	tsvector	Atlas Search	Nativo
Conectividade (Python)	Sim	Sim	Sim

Fonte: Dados da pesquisa.

Ao observar as características dos bancos de dados, o *software* escolhido foi o *Elasticsearch* pelas características técnicas nativas, principalmente nos tópicos que atuam diretamente sobre o processo de FTS. O *Elasticsearch* é uma solução composta de muitos recursos, dentre o quais destaca-se o banco de dados distribuído e orientado a documentos, as ferramentas de gestão e análise de dados, a captação e importação, dentre outros. Optou-se pela sua instalação em *Docker*[39], um *software* auxiliar que permite o acesso a outros *softwares* por um conceito denominado "*container*". O *Docker* fornece uma camada de abstração e virtualização nos principais sistemas operacionais e permite a execução dos programas pelo carregamento de suas imagens.

Um mapeamento do esquema do banco de dados foi adotado pelo autor para padronizar os metadados escolhidos e melhorar a descrição de cada artigo armazenado, na

[38] Full text search
[39] https://docs.docker.com/get-started/

hipótese de facilitar a interoperabilidade futura. No *Elasticsearch* esta estrutura é denominada *índice*. Nesse projeto foi usado o seguinte mapeamento baseado em metadados Dublin Core[40]:

Quadro 7 - Estrutura de índice para o Elasticsearch

```
PUT articles
{
 "mappings": {
  "properties": {
   "dcIdentifier": {
    "type": "text"
   },
   "dcDate": {
    "type": "text"
   },
   "dcLanguage": {
    "type": "text"
   },
   "dcCreator": {
    "type": "text"
   },
   "dcTitle": {
    "type": "text"
   },
   "dcDescription": {
    "type": "text"
   },
   "dcSubject": {
    "type": "text"
   },
   "dcSource": {
    "type": "text"
   },
   "dcFormat": {
    "type": "text"
   },
```

[40] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

```
    "keywords": {
      "type": "text"
    },
    "textBody": {
      "type": "text"
    }
  }
 }
}
```

Fonte: Dados da pesquisa

Para a captação da estrutura de extração do texto do artigo, desenvolveu-se um algoritmo em linguagem Python à parte do desenvolvimento do *software* principal, onde a biblioteca de extração de pdf, forneceu os resultados em dois grandes conjuntos de dados em uma lista: na primeira posição há o corpo do texto do pdf, na segunda posição há um dicionário de dados com a estrutura "chave : valor", com a qual pode-se acessar os metadados. Para facilitar o entendimento, segue uma correspondência entre o *Schema* definido no banco de dados e os dados extraídos do arquivo pdf (quadro 8):

Quadro 8 - Correspondência entre o esquema do banco de dados e dados extraídos do pdf

Schema no Elasticsearch	**Texto e Metadados do pdf**	**Descrição**
dcIdentifier	"Hash SHA1"	Código gerado pelo algoritmo Python
dcDate	creationDate	Data de criação do artigo
dcLanguage	"Fixo"	Como a biblioteca não possui o metadado, usou-se "en" para inglês
dcCreator	author	Autor[41](*)
dcTitle	title	Título
dcDescription	-	Descrição do artigo, na biblioteca Tika
dcSubject	subject	Assunto
dcSource	"Nome do arquivo"	Nome do arquivo extraído do caminho completo
dcFormat	format	Formato da codificação do arquivo

[41] (*) Na biblioteca Tika, esse metadado é um *array* para vários autores; na biblioteca pyMuPDF, é uma *string* apenas com o primeiro autor

keywords	keywords	Palavras chave
textBody	“Corpo do artigo”	Texto do artigo científico

Fonte: Dados da pesquisa.

Importante notar que os metadados dependem do preenchimento em sua origem, ou seja, na preparação do artigo na época em que ele foi criado. Dessa forma, um percentual significativo dos artigos não possuía metadados, ainda que os mais recentes possuíssem um conjunto expressivo para consulta e utilização. Nesse projeto os metadados atuam como uma possibilidade adicional na identificação do artigo recuperado, mas a pesquisa enfatiza o texto que compõe o corpo do artigo e seu título.

Antes de popular o banco de dados com os artigos sob análise, o *software* (*Elasticsearch*) foi instalado e configurado, um esquema foi definido, e o algoritmo (Python) foi criado. Esse algoritmo realiza a carga do banco de dados e obedeceu a seguinte ordem:

1. Abre-se uma conexão com *Elasticsearch*, fazendo referência ao índice desejado;
2. Prepara-se uma lista de todos arquivos pdf a serem processados;
3. Inicia-se uma iteração da lista:
    3.1. Calcula-se o código *Hash* do arquivo pdf;
    3.2. Submete-se consulta para verificar se o artigo já foi gravado: se já foi gravado, processa-se o próximo da lista; se ainda não foi, segue-se o algoritmo;
    3.3. Extrai-se o texto e os metadados do artigo;
    3.4. Efetua-se o pré-processamento no corpo textual com as seguintes etapas:
        3.4.1. Retirar *stopwords*[42];
        3.4.2. Retirar quebras de linha;
        3.4.3. Retirar hifens;
        3.4.4. Tratar texto em *Unicode*;
        3.4.5. Transformar o texto em minúsculas;
        3.4.6. Retirar URLs;
        3.4.7. Retirar endereços de *e-mail*;
        3.4.8. Retirar números de telefone;
        3.4.9. Retirar símbolos monetários;
        3.4.10. Retirar pontuação;
        3.4.11. Retirar caracteres especiais;

[42] Processo parcial como visto anteriormente

3.4.12. Padronizar espaços extras entre palavras;

3.5. Cria-se o conjunto de dados em formato JSON contendo todos os dados necessários para o armazenamento do artigo no banco de dados;

3.6. O conjunto de dados JSON é enviado para o banco de dados e repete-se o processo para o próximo arquivo pdf da lista.

Após esse processo, considera-se o banco de dados de artigos carregado. Pode-se então submeter consultas ao banco de dados, o qual, por sua vez, retorna um conjunto de dados em formato JSON para interação com a interface do usuário.

5.2.2.2 Os passos do experimento em si

Um importante motivador da pesquisa, era entender e identificar como os artefatos terminológicos da área médica poderiam melhorar o processo de RI. A estratégia foi uma revisão de literatura acompanhada da construção de um software para a RI de artigos científicos mediante a influência das terminologias clínicas. A presente seção retrata, portanto, os passos realizados na construção do algoritmo, bem como as tecnologias e estratégias usadas. A figura 4 é uma referência gráfica dos tópicos apresentados a seguir:

1. Seleção de termos comuns – MeSH;
2. Submissão de consultas – MeSH;
   2.1. Seleção dos descritores hierárquicos – MeSH;
   2.2. Seleção de termos de entrada – MeSH;
3. Submissão de consultas (MeSH) usando expansão;
4. Submissão de consultas – SNOMED CT;
   4.1. Seleção de identificador – SNOMED CT;
   4.2. Seleção de termos hierárquicos – SNOMED CT;
   4.3. Seleção de termos conceitualmente próximos – SNOMED CT;
5. Submissão de consultas (SNOMED CT) usando expansão.

<u>1. Seleção de termos comuns - MeSH</u>

As duas terminologias selecionadas contêm milhares de termos relacionados à área médica. A fim de permitir um conjunto de dados compatível com a temática do acervo de artigos, estabeleceu-se um recorte apropriado para análise. Observou-se que a terminologia MeSH está organizada em 16 grandes grupos temáticos, a saber:

1. *Anatomy [A]*
2. *Organisms [B]*
3. *Diseases [C]*
4. *Chemicals and Drugs [D]*
5. *Analytical, Diagnostic and Therapeutic Techniques, and Equipment [E]*
6. *Psychiatry and Psychology [F]*
7. *Phenomena and Processes [G]*
8. *Disciplines and Occupations [H]*
9. *Anthropology, Education, Sociology, and Social Phenomena [I]*
10. *Technology, Industry, and Agriculture [J]*
11. *Humanities [K]*
12. *Information Science [L]*
13. *Named Groups [M]*
14. *Health Care [N]*
15. *Publication Characteristics [V]*
16. *Geographicals [Z]*

Dessa forma o grupo [C] *Diseases*, foi escolhido como base para comparação entre as estruturas terminológicas. O algoritmo então realizou as seguintes etapas:

1. Seleção dos termos a partir da MeSH iniciados com o código “C”;
2. Para cada termo selecionado, verifica-se a correspondência na SNOMED CT;
3. O termo que encontra correspondência é organizado em uma lista em memória;

A lista resultante desse processo foi usada como entrada do processo de pesquisa. Para cada termo encontrado em comum entre as terminologias, foram feitas consultas com os respectivos termos hierárquicos e relacionados. A fim de ilustrar os principais processos na forma em que ocorrem no algoritmo, a figura 4 exibe os diversos passos do experimento, procurando demonstrar a entrada, processamento e saída.

Cada terminologia possui particularidades em estrutura e organização de conceitos, de modo que suas atividades são detalhadas separadamente: primeiro a MeSH e depois a SNOMED.

2. Submissão de consultas - MeSH

Esta função de pesquisa na terminologia MeSH é invocada como uma *thread*[43], definida por dois argumentos do tipo *string*: o *termo*[44] e o *indicador de idioma*[45]. Para cada termo comum selecionado, o algoritmo fez uma consulta no banco de dados com a MeSH e identificará o termo específico na terminologia – nesse caso um descritor – e o respectivo código (*descriptor UI)* associado.

O *descriptor UI* é obtido por uma função que recebe três argumentos com os seguintes parâmetros: i) o termo procurado, ii) o tipo do termo, se original ou tratado, iii) idioma. A resposta retorna um código alfanumérico, por exemplo, *D009203*. Caso o termo procurado seja um *termo de entrada* o algoritmo retorna o código do descritor principal e sua descrição textual, conforme exemplo da Figura 4. A partir desse momento o *descriptor UI* foi usado para que outras funções submetam consultas ao banco de dados. A continuidade desse processo tem como sequência dois grandes processos, os quais são analisados em seguida: i) a seleção dos descritores hierárquicos e ii) a seleção dos termos de entrada de cada descritor.

2.1. Seleção dos descritores hierárquicos - MeSH

A seleção de descritores hierárquicos retorna um conjunto de termos, também chamados "cabeçalhos" na MeSH. Em sua estrutura é possível notar a importância dos descritores enquanto referências relacionadas ao conceito. Cada descritor possui um identificador único, em código, conforme visto, caracterizado pelo *descriptor UI*. Como um descritor pode possuir vários termos de entrada, o identificador é somente do descritor principal, os termos de entrada não possuem códigos identificadores. De modo que, se o usuário solicitar uma consulta usando um termo de entrada, o algoritmo localiza o descritor agregador (preferido) desse termo. Para recuperar os descritores hierárquicos, o código recebe, novamente, os três argumentos: i) termo procurado, ii) tipo do termo, se original ou tratado, iii) idioma. Em seguida, os seguintes passos são realizados:

- Selecionam-se os códigos hierárquicos a partir do *descriptor UI* do descritor;
- Identificam-se os respectivos descritores textuais a partir de cada código hierárquico;

[43] É uma tarefa organizada pelo sistema operacional que pode solicitar à CPU tarefas simultâneas em processadores com mais de um núcleo, ou alternadas, gerando um efeito de paralelismo.
[44] Os termos foram tratados anteriormente para padronização em minúsculas (*case folding*).
[45] Visto que as terminologias possuem portabilidade para outros idiomas, o algoritmo já foi desenvolvido para utilização futura em *multi-language*.

- Cria-se uma lista de descritores (sem repetição) a partir de cada grupo hierárquico;
- Retorna-se uma lista de descritores ao código de chamada.

A exemplo da figura 4, observa-se a ilustração para demonstrar o processo: a partir da pesquisa do termo "*heart attack*" identifica-se o "*descriptor UI*: *D009203*" e o descritor principal (preferido) "*myocardial infarction*".

Invocando a função para retornar os códigos hierárquicos, passando como argumentos o identificador "*D009203*" e o idioma "eng", obtém-se respectivamente "C14280647500", "C14907585500", "C23550513355750" e "C23550717489750". Dessa lista de códigos hierárquicos, realiza-se uma interação onde outra função retorna os descritores alinhados a esta hierarquia, em forma de uma lista de descritores. Esses descritores estão hierarquicamente relacionados ao descritor preferido – *myocardial infarction* – nesse exemplo: *non-st elevated myocardial infarction*, *inferior wall myocardial infarction*, *anterior wall myocardial infarction*, *shock, cardiogenic* e *st elevation myocardial infarction*. Esse processo, totaliza assim seis descritores hierárquicos relacionados ao termo de código *D009203*.

2.2. Seleção de termos de entrada - MeSH

A seleção de termos de entrada considera que, consequentemente, para cada descritor recupera-se os termos de entrada. Tais termos são importante representação conceitual, e muitas vezes estão mais próximos da linguagem natural, expressa pelo usuário do que pelo descritor principal, que pode se aproximar de uma linguagem mais técnica.

Uma outra função, de forma bastante similar à função de descritores hierárquicos, recebe três argumentos: i) descritor principal, ii) tipo do termo (se original ou tratado), iii) idioma. Desse modo, uma lista com os termos de entrada, relacionados aos descritores será criada da seguinte forma:

i. Identifica-se o código do descritor principal
ii. Encontra-se uma lista de códigos de descritores hierárquicos, relacionados ao descritor principal
iii. Para cada item dessa lista, identifica-se os códigos dos descritores relacionados
iv. Para cada código relacionado, retornar os termos de entrada daquele descritor.

3. Submissão de consultas (MeSH) usando expansão

Com as listas de descritores e termos de entrada prontas na memória, o processo avança para a etapa de consulta ao banco de dados de artigos, com as seguintes tarefas:

i. Uma conexão com o banco de dados de artigos é aberta, e o índice é referenciado;
ii. Para cada descritor e termo de entrada encontrado, submete-se uma consulta ao banco de dados.

Utilizando a biblioteca Python para *Elasticsearch* 7.7.0, a definição da consulta tem a seguinte sintaxe, onde "desc" representa o termo descritor:

Quadro 9 - Modelo de consulta no Elasticsearch

```
"consulta": {
 "multi_match": {
  "consulta": desc,
  "type": "phrase",
  "fields": [
   "dcTitle",
   "textBody"
  ]
 }
}
```

Fonte: Dados da pesquisa.

Caso exista revocação, adiciona-se um conjunto de dados em uma lista em memória, qual contém:

- Termo descritor pesquisado;
- dcIdentifier ou o *hash* do artigo;
- dcSource ou nome do arquivo do artigo pdf;
- dcTitle ou o título do artigo (quando disponível).

A figura 4, exibe esse processo e o conjunto de dados resultante é parte de um conjunto maior com a adição das respostas da pesquisa aos termos da outra terminologia.

4. Submissão de consultas com a SNOMED CT

O processo de pesquisa usando a SNOMED CT é similar à etapa 2, realizado com a MeSH, mas existem particularidades que derivam no fato das terminologias possuírem estruturas diferentes, mesmo mantendo em comum uma hierarquia. Nesse trabalho, utilizaram-se dois grandes grupos a partir de conceitos disponíveis na terminologia: os descritores

hierárquicos e os descritores conceitualmente próximos. Não há na SNOMED, o "termo de entrada" como no MeSH. Uma visão específica do processo de pesquisa com a terminologia SNOMED CT também pode ser vista na figura 4.

4.1. Seleção de identificador – SNOMED CT

Da mesma forma que o processo com a MeSH, a submissão da consulta através da SNOMED CT, inicia-se com o recebimento do termo (*string*) a ser pesquisado. Há um código identificador para cada conceito, porém com uma diferença: um código, equivalente a um conceito, pode possuir várias descrições (vide Seção 3.1.3). Há uma descrição principal e as demais descrições são consideradas por proximidade conceitual. A título de exemplo, a partir da mesma pesquisa submetida à MeSH, com o termo *heart attack*, abre-se uma conexão com o banco de dados do SNOMED CT para selecionar o identificador principal do termo procurado. Para esta identificação são realizadas as tarefas (figura 4):

- Recebe-se termo para pesquisa, seu tipo e idioma;
- Abre-se uma conexão com a terminologia SNOMED CT no banco;
- Selecionam-se os identificadores (conceitos) relacionados ao termo procurado
- Seleciona-se o conceito principal do conjunto de identificadores, no exemplo, "*22298006*".

4.2. Seleção de termos hierárquicos – SNOMED CT

Após a identificação do código que identifica o conceito, outra função para seleção de termos hierárquicos recebe três argumentos: i) o identificador principal, ii) tipo do termo, se original ou tratado, iii) idioma. Considerando uma visão a partir do início da solicitação do usuário, pode-se considerar que a seleção possui os seguintes passos:

- Seleciona os termos hierárquicos, a partir do identificador principal (*22298006);*
- Cria uma lista de descritores (sem repetição) a partir de cada grupo hierárquico;
- Retorna a lista de descritores (*array*) ao código que o invocou.

A partir do código de identificação do conceito é possível selecionar os termos relacionados hierarquicamente. Os termos hierárquicos são selecionados a partir do identificador principal, usado para selecionar os identificadores relacionados àquele código e, consequentemente, recuperar suas descrições textuais.

4.3. Seleção de termos sinônimos – SNOMED CT

Embora a SNOMED, em comparação com a MeSH, não contenha “termos de entrada”, é possível traçar um paralelo considerando termos sinônimos, os quais respondem ao mesmo identificador do conceito. Para recuperar tais termos, outra função é invocada, a qual retorna uma lista de termos em memória.

5. Submissão de consultas (SNOMED CT) usando expansão

De posse de ambas listas – termos hierárquicos e termos sinônimos – uma conexão com o banco de dados no *ElasticSearch* é aberta para o índice dos artigos. Inicia-se um conjunto de comandos que faz interagir ambas as listas, começando pelos termos hierárquicos e depois passando aos sinônimos. Para cada termo submete-se uma consulta:

Quadro 10 - Modelo de consulta no Elasticsearch

```
"query": {
 "multi_match": {
  "consulta": termoHierq,
  "type": "phrase",
  "fields": [
   "dcTitle",
   "textBody"
  ]
 }
}
```

Fonte: Dados da pesquisa.

Para uma melhor explicação, uma ilustração do processo pode ser vista na figura 4. Para cada resposta do banco, um conjunto de dados é selecionado entre os diversos metadados para serem armazenados:

- Termo descritor pesquisado;
- dcIdentifier ou *Hash* do artigo;
- dcSource ou nome do arquivo do artigo pdf;
- dcTitle ou título do artigo (quando disponível).

Esta estrutura é a mesma apresentada para MeSH a fim de estabelecer um padrão de análise posterior de ambas terminologias clínicas. Ao final desse processo, uma lista de respostas é codificada em formato JSON e retorna para o código inicial, além de ser gravado

em arquivo para posterior consulta e conferência. Os resultados do processo de submissão de consultas são apresentados no capítulo 6.

6 RESULTADOS E DISCUSSÃO

Nesta seção são discutidos os resultados obtidos nas etapas da metodologia. Espera-se finalmente evidenciar as conclusões de cada etapa e demonstrar a contribuição para o experimento em seu objetivo completo.

6.1 Resultados

A presente seção está organizada para refletir as seções equivalentes do Capítulo 5, especificamente, as seções 5.2.2.1 e 5.2.2.2.

6.1.1 Resultados dos passos iniciais

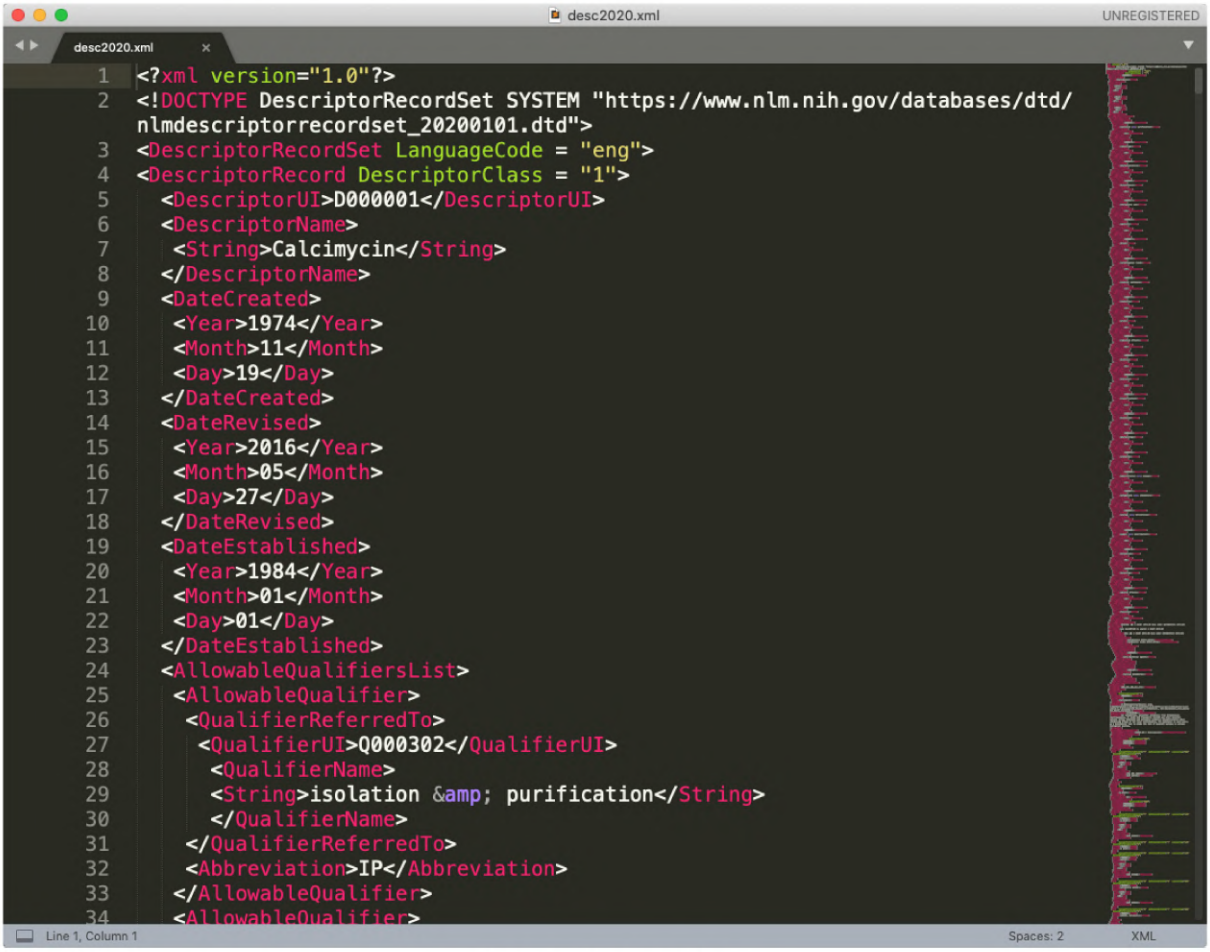
6.1.1.1 Resultado da aquisição das terminologias

A aquisição das terminologias resultou em dois artefatos, arquivos eletrônicos contendo as terminologias. Os arquivos possuem formatos distintos para a publicação de dados. O modo de publicação da MeSH é codificado em XML. Conforme citado na metodologia (seção 5.2.2.1), foi criado um algoritmo para percorrer o arquivo usando a *ElementTree*[46] como biblioteca principal de iteração para extração dos dados.

A codificação original da MeSH possui uma grande quantidade de metadados em tags. Juntamente com o XML, há um arquivo *Document Type Definition* (DTD) que contém as regras de estrutura de elementos e atributos XML. Não foram usados todos os elementos no presente projeto. Para que o experimento fosse implementado, o arquivo XML foi transformado em um banco de dados relacional. A figura 8 demonstra como resultado da etapa de aquisição da MeSH, seu arquivo codificado.

[46] https://lxml.de/tutorial.html

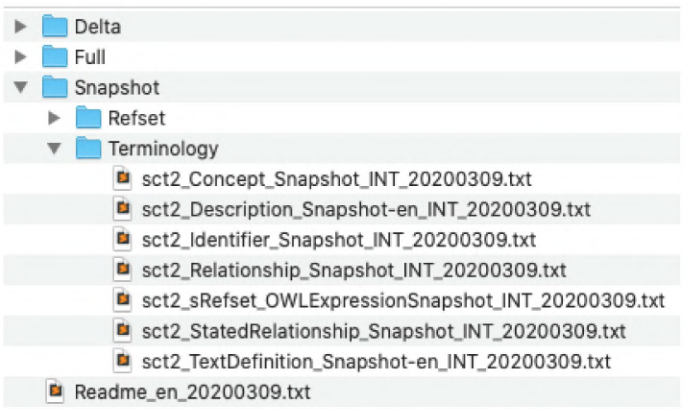
Figura 7 - Resultado da aquisição da terminologia MeSH



Fonte: Dados da pesquisa

A SNOMED CT é distribuída em arquivo compactado no formato .zip que ao ser descompactado exibe uma estrutura de pastas e arquivos no sistema operacional, como exemplificado na figura 9.

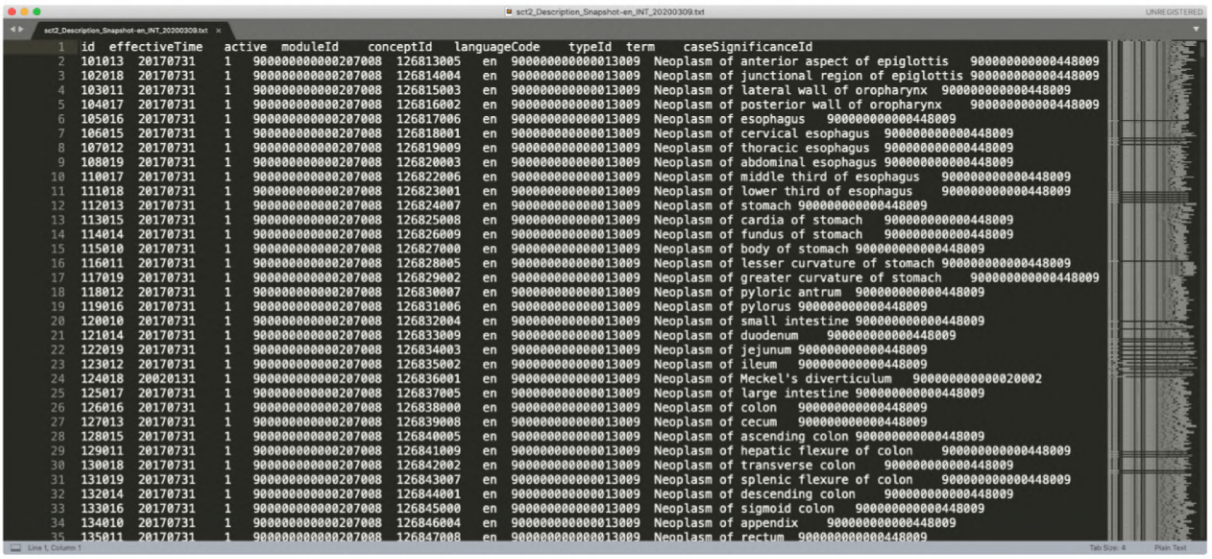
Figura 8 - Arquivos como resultado da aquisição da SNOMED CT



Fonte: Dados da pesquisa

Os arquivos estão em formato TXT, sendo que a primeira linha contém o descritivo da coluna, ou seja, cada linha é um registro cujos dados estão separados por tabulação TSV, (Figura 10). A estrutura é praticamente um modelo relacional, sendo a coluna ID, em quase todos arquivos, a chave primária da tabela. Após análise dos arquivos e entendimento pela documentação da MLDS, conforme citado na metodologia (seção 5.2.2.1), foi realizado a migração do formato para um formato relacional. A figura 10 demonstra o conteúdo de um arquivo adquirido da SNOMED.

Figura 9 - Conteúdo de arquivo SNOMED CT



id	effectiveTime	active	moduleId	conceptId	languageCode	typeId	term	caseSignificanceId
101013	20170731	1	900000000000207008	126813005	en	900000000000013009	Neoplasm of anterior aspect of epiglottis	900000000000448009
102018	20170731	1	900000000000207008	126814004	en	900000000000013009	Neoplasm of junctional region of epiglottis	900000000000448009
103011	20170731	1	900000000000207008	126815003	en	900000000000013009	Neoplasm of lateral wall of oropharynx	900000000000448009
104017	20170731	1	900000000000207008	126816002	en	900000000000013009	Neoplasm of posterior wall of oropharynx	900000000000448009
105016	20170731	1	900000000000207008	126817006	en	900000000000013009	Neoplasm of esophagus	900000000000448009
106015	20170731	1	900000000000207008	126818001	en	900000000000013009	Neoplasm of cervical esophagus	900000000000448009
107012	20170731	1	900000000000207008	126819009	en	900000000000013009	Neoplasm of thoracic esophagus	900000000000448009
108019	20170731	1	900000000000207008	126820003	en	900000000000013009	Neoplasm of abdominal esophagus	900000000000448009
110017	20170731	1	900000000000207008	126822006	en	900000000000013009	Neoplasm of middle third of esophagus	900000000000448009
111018	20170731	1	900000000000207008	126823001	en	900000000000013009	Neoplasm of lower third of esophagus	900000000000448009
112013	20170731	1	900000000000207008	126824007	en	900000000000013009	Neoplasm of stomach	900000000000448009
113015	20170731	1	900000000000207008	126825008	en	900000000000013009	Neoplasm of cardia of stomach	900000000000448009
114014	20170731	1	900000000000207008	126826009	en	900000000000013009	Neoplasm of fundus of stomach	900000000000448009
115010	20170731	1	900000000000207008	126827000	en	900000000000013009	Neoplasm of body of stomach	900000000000448009
116011	20170731	1	900000000000207008	126828005	en	900000000000013009	Neoplasm of lesser curvature of stomach	900000000000448009
117019	20170731	1	900000000000207008	126829002	en	900000000000013009	Neoplasm of greater curvature of stomach	900000000000448009
118012	20170731	1	900000000000207008	126830007	en	900000000000013009	Neoplasm of pyloric antrum	900000000000448009
119016	20170731	1	900000000000207008	126831006	en	900000000000013009	Neoplasm of pylorus	900000000000448009
120010	20170731	1	900000000000207008	126832004	en	900000000000013009	Neoplasm of small intestine	900000000000448009
121014	20170731	1	900000000000207008	126833009	en	900000000000013009	Neoplasm of duodenum	900000000000448009
122019	20170731	1	900000000000207008	126834003	en	900000000000013009	Neoplasm of jejunum	900000000000448009
123012	20170731	1	900000000000207008	126835002	en	900000000000013009	Neoplasm of ileum	900000000000448009
124018	20020131	1	900000000000207008	126836001	en	900000000000013009	Neoplasm of Meckel's diverticulum	900000000000020002
125017	20170731	1	900000000000207008	126837005	en	900000000000013009	Neoplasm of large intestine	900000000000448009
126016	20170731	1	900000000000207008	126838000	en	900000000000013009	Neoplasm of colon	900000000000448009
127013	20170731	1	900000000000207008	126839008	en	900000000000013009	Neoplasm of cecum	900000000000448009
128015	20170731	1	900000000000207008	126840005	en	900000000000013009	Neoplasm of ascending colon	900000000000448009
129011	20170731	1	900000000000207008	126841009	en	900000000000013009	Neoplasm of hepatic flexure of colon	900000000000448009
130018	20170731	1	900000000000207008	126842002	en	900000000000013009	Neoplasm of transverse colon	900000000000448009
131019	20170731	1	900000000000207008	126843007	en	900000000000013009	Neoplasm of splenic flexure of colon	900000000000448009
132014	20170731	1	900000000000207008	126844001	en	900000000000013009	Neoplasm of descending colon	900000000000448009
133016	20170731	1	900000000000207008	126845000	en	900000000000013009	Neoplasm of sigmoid colon	900000000000448009
134010	20170731	1	900000000000207008	126846004	en	900000000000013009	Neoplasm of appendix	900000000000448009
135011	20170731	1	900000000000207008	126847008	en	900000000000013009	Neoplasm of rectum	900000000000448009

Fonte: Dados da pesquisa

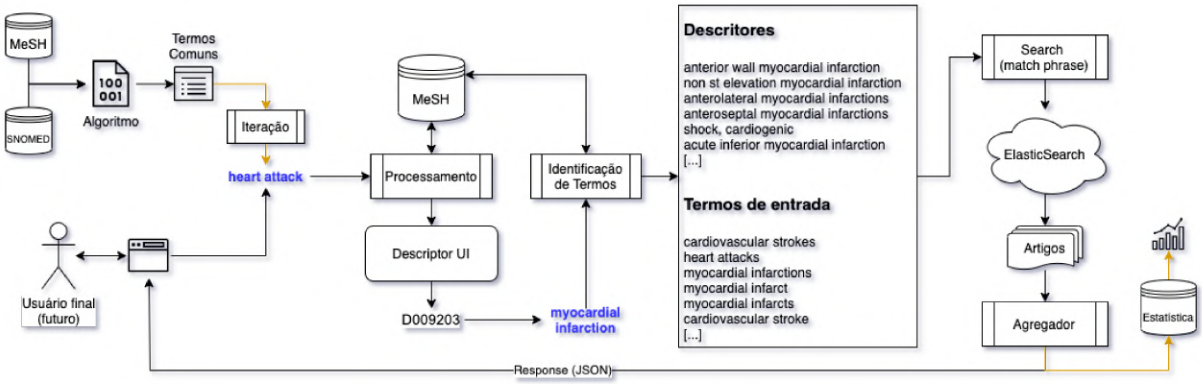
6.1.1.2 Resultado da estruturação terminológica

A estruturação terminológica resultou em dois bancos de dados para a manipulação das terminologias em um modelo ER, a fim de permitir o acesso às terminologias por linguagem SQL na consulta de termos e definições hierárquicas. Tal processo foi mais simples na MeSH devido sua concepção que já utiliza códigos para a identificação única dos termos e sua posição de relações hierárquicas.

A estrutura da MeSH estabelece como princípio, de forma a proporcionar hierarquias múltiplas para certo descritor, que o descritor é representado uma vez e possui um código de identificação (chave) associado a ele. Existe um ou mais de um código (*TreeNumberList*) que permitem associar o descritor a diferentes locais na hierarquia. Consequentemente, são possíveis relações múltiplas que confere, flexibilidade ao resultado final. A figura 3 exemplifica o descritor "*Myocardial Infarction*" que possui quatro elementos

"*TreeNumberList*", possibilitando que ele surja em 4 posições distintas na terminologia. Abaixo, na figura 11, pode-se observar o processo mais amplo de identificação de termos na MeSH.

Figura 10 - O descritor principal e seus desdobramentos na MeSH



Fonte: Dados da pesquisa

Já os termos conceitualmente próximos, os quais direcionam o usuário a usar o descritor principal, são elencados por uma lista no arquivo de origem pela tag "<TermList>". A figura 11 também permite visualizar que, para o descritor "*Myocardial Infarction*", existem termos de entrada (*Entry Terms*), a saber: "*cardiovascular strokes*", "*heart attack*", "*myocardial infarction*", "*myocardial infarct*", "*myocardial infarcts*", "*cardiovascular strokes*", entre outros, totalizando 35 termos de entrada relacionados à hierarquia do descritor principal. *Cardiovascular Stroke*, *Heart Attack, Myocardial Infarct.* A estruturação da MeSH resultou, portanto, em um banco composto por três tabelas:

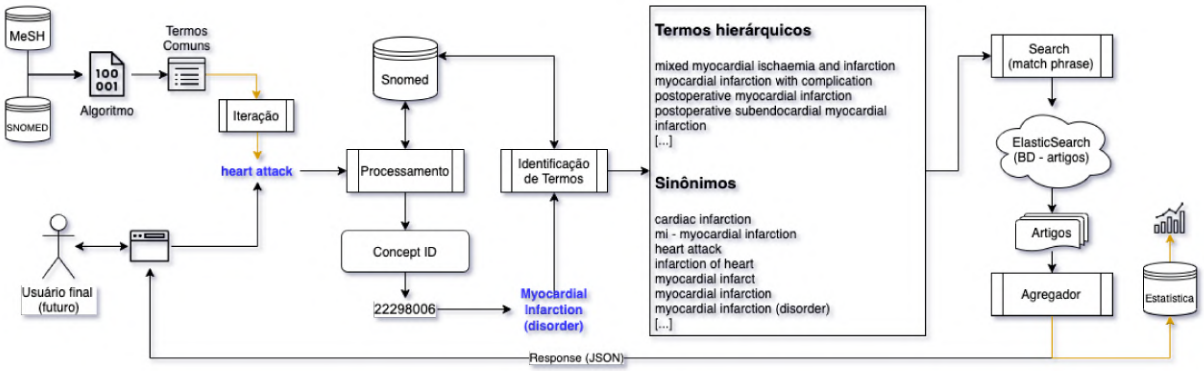
- Descritores, que armazena os descritores principais (preferidos);
- Termos, que armazena os termos de entrada (descritores não-preferidos);
- Hierarquia, que armazena os códigos hierárquicos.

Esse banco de dados é usado no experimento, a fim de prover as informações da terminologia de modo estruturado e em tempo de processamento razoável para o processo de expansão das consultas.

Na *SNOMED CT* a estruturação foi mais complexa. Embora a distribuição assemelhar-se a um modelo relacional, a estrutura hierárquica não segue um modelo de codificação semelhante ao MeSH: as propriedades que definem a hierarquia estão gravadas em axiomas da linguagem OWL. Além disso, há dois modelos de interação com a terminologia: *stated* e *inferred,* ou seja, declarado e inferido. Adotou-se a forma *declarada* para identificação dos termos, uma vez que o inferido utiliza uma *reasoner* para gerar novas declarações.

Com relação aos termos hierárquicos, a tabela *refset* é a estrutura que armazena as declarações em axiomas e que, assim, permite a identificação da estrutura hierárquica. De posse do código de descritor específico, por exemplo "*myocardial infarction (disorder)*", cujo identificador é 22298006 (Figura 4), o código realiza uma chamada recursiva a cada identificação de relação. Isso permite identificar as subclasses dos identificadores selecionados repetindo o algoritmo. Com esta estratégia, obtém-se a árvore terminológica, hierárquica, dado o termo consultado. A figura 12 também permite observar os demais termos associados a partir do *id* principal.

Figura 11 - O conceito principal e seus desdobramentos na SNOMED CT



Fonte: Dados da pesquisa

Os termos sinônimos estão na figura em destaque (figura 12), a partir do descritor "*Myocardial infarction (disorder)*". Nesse exemplo:

- *Cardiac infarction*
- *MI - myocardial infarction*
- *Heart attack*
- *Infarction of heart*
- *Myocardial infarct*

O algoritmo para localizar esses termos usa o identificador do termo principal "22298006" e localiza na tabela "*description*" onde estão armazenadas as descrições, ou seja, os textos relacionados a esse identificador específico. Nesse exemplo, os termos hierárquicos recuperados foram os seguintes: "*acute myocardial infarction with rupture of ventricle*", "*acute myocardial infarction with rupture of ventricle (disorder)*", "*mixed myocardial ischaemia and infarction*", "*mixed myocardial ischemia and infarction*", "*mixed myocardial ischemia and infarction (disorder)*", "*myocardial infarction with complication*", "*myocardial infarction with*

complication (disorder)", "*postoperative myocardial infarction*", "*postoperative myocardial infarction (disorder)*", "*postoperative subendocardial myocardial infarction*", "*postoperative subendocardial myocardial infarction (disorder)*", "*postoperative transmural myocardial infarction of anterior wall*", "*postoperative transmural myocardial infarction of anterior wall (disorder)*", "*postoperative transmural myocardial infarction of inferior wall*", "*postoperative transmural myocardial infarction of inferior wall (disorder)*".

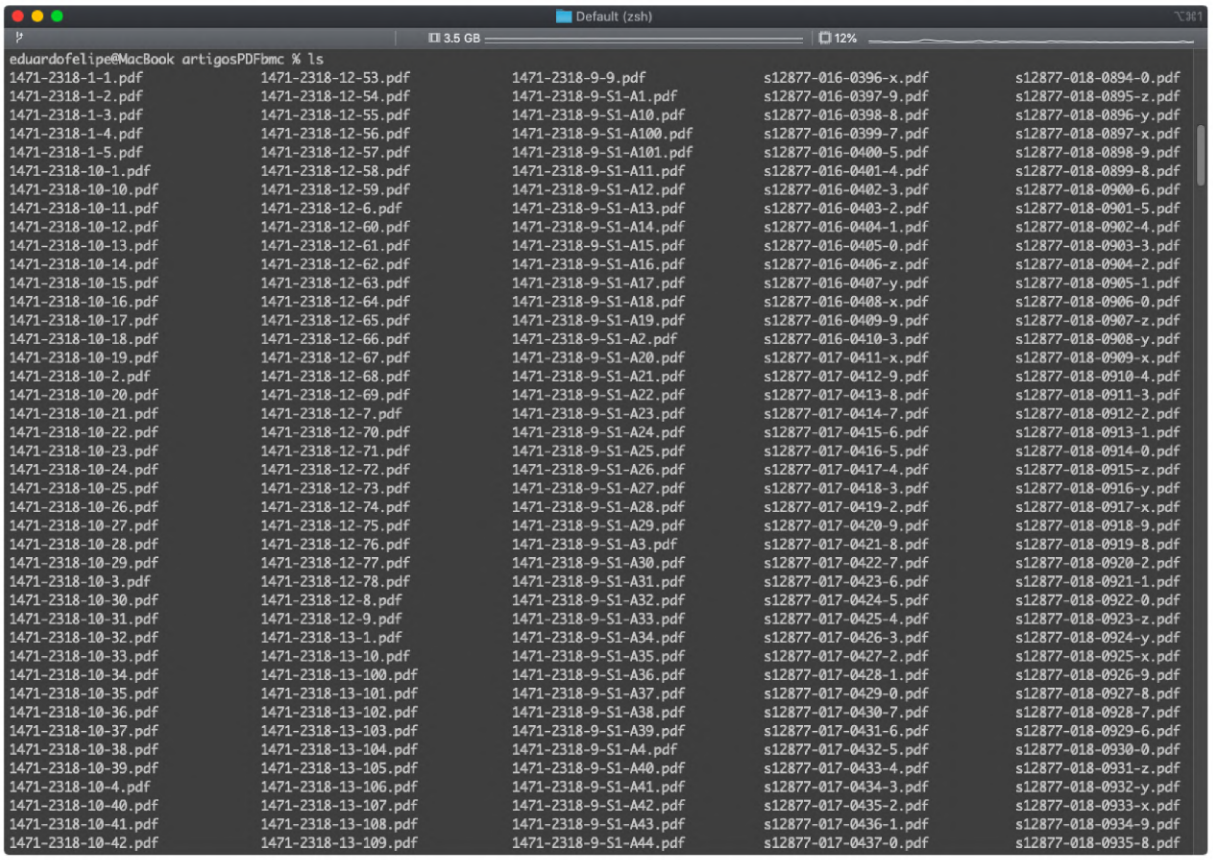
Na estruturação da terminologia SNOMED CT, quatro tabelas foram mapeadas:

- *Concept*, que armazena os códigos de conceitos
- *Description*, que armazena as descrições
- *Statedrelationship*, que armazena as referências de origem e destino dos identificadores
- *Refset*, que armazena os axiomas

6.1.1.3 Resultado da aquisição de artigos

A aquisição dos artigos para o banco de dados foi uma tarefa crítica. Afinal, as consultas foram submetidas para localizar textos desses artigos. Os critérios para a escolha (seção 5.2.2.1), observaram a necessidade em trabalhar com artigos completos e não apenas *abstracts*. Após a escolha da instituição dos artigos, e a certeza da disponibilidade dos documentos em texto completo, desenvolveu-se o algoritmo para *download* dos arquivos de forma automatizada. Como resultado desse processo de *scraping*, os arquivos coletados totalizaram 2.481 artigos, e foram armazenados em uma pasta no computador local, conforme a Figura 13 abaixo:

Figura 12 - Fragmento do resultado do processo de scraping



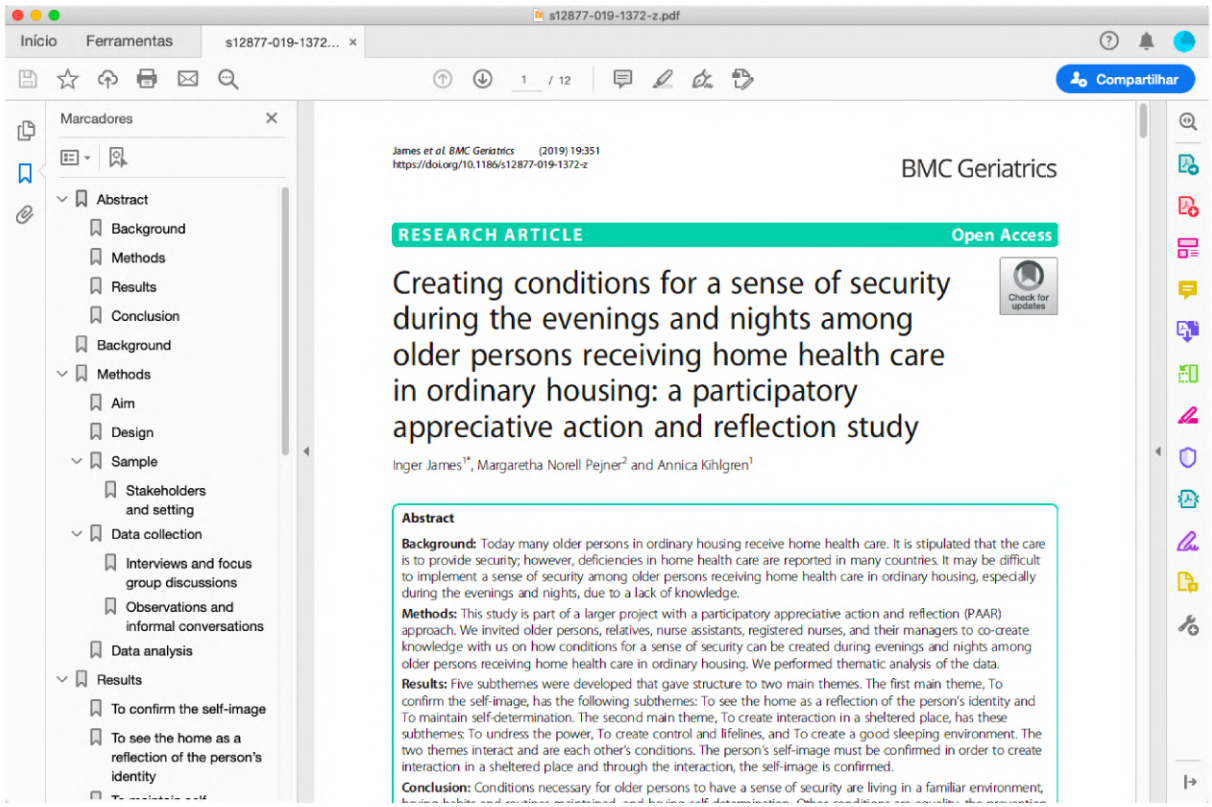
Fonte: Dados da pesquisa

A partir do conjunto de artigos completos em pdf, foi realizado a extração de textos e sua preparação, de forma a constituir o conjunto de dados referência para as consultas expandidas.

6.1.1.4 Resultado do pré-processamento

O pré-processamento é um processo crucial para o experimento. A adequação dos dados para o processo de recuperação é necessária visto o formato proprietário dos arquivos publicados em pdf. Os arquivos passaram por uma série de etapas, desde seu *download*, extração e tratamento. A Figura 14 mostra o formato inicial quando é feito o download do artigo pdf.

Figura 13 - Formato original do arquivo pdf

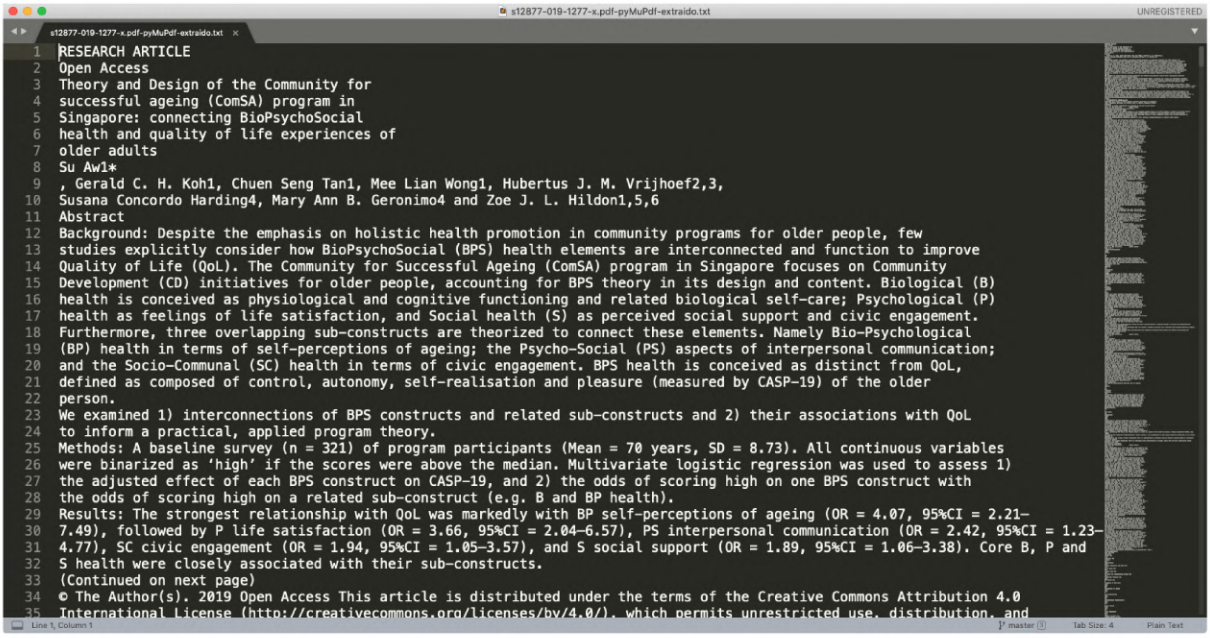


Fonte: Dados da pesquisa

Na primeira etapa de pré-processamento, a extração textual possibilita o acesso ao texto puro (ANSI), sem adição de codificação e formatação por padrões proprietários, como pdf, docx, etc. Esse processo foi realizado por um algoritmo criado em linguagem Python que usou a biblioteca *PyMuPDF*, conforme descrito na seção 5.2.21. O algoritmo do experimento foi escrito para aceitar duas bibliotecas na extração (*PyMuPDF* e *Apache Tika-Python*), com a possibilidade de configuração de escolha caso fosse necessário.

Observa-se ainda (Figura 15), que o texto extraído do pdf contém quebras de linhas, além de pontuação, letras maiúsculas, caracteres especiais entre outros, aproximando-se da representação do arquivo original. Embora esse formato já esteja em texto, é necessário o tratamento, denominado pré-processamento.

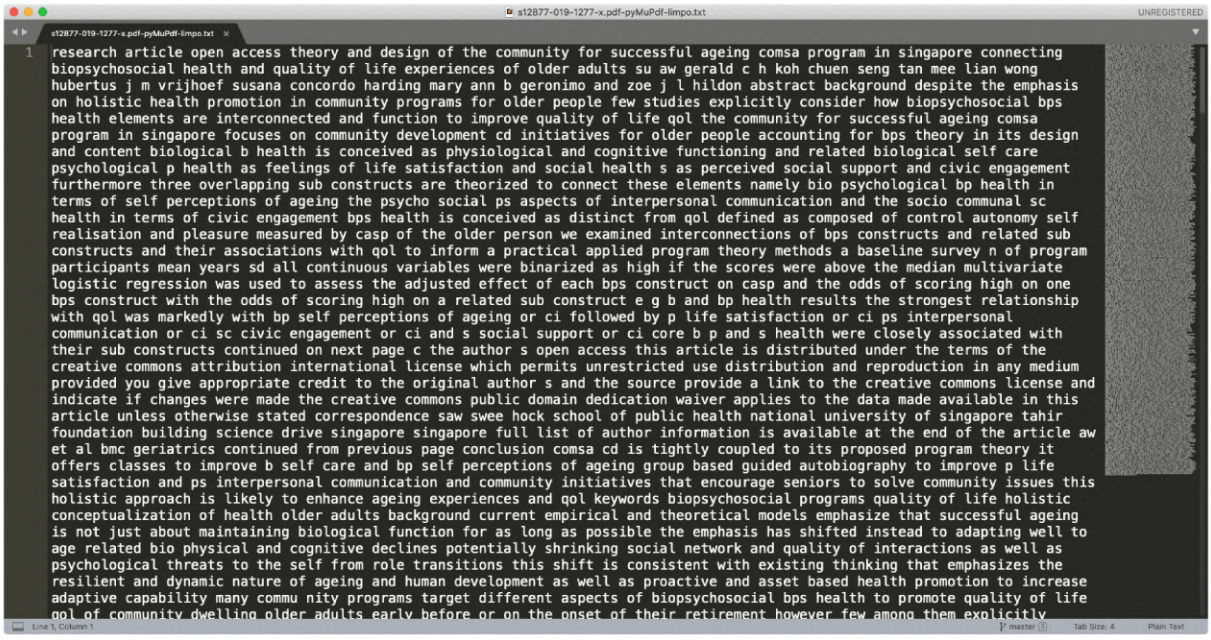
Figura 14 - Texto extraído do arquivo pdf sem tratamento



Fonte: Dados da pesquisa

Após o pré-processamento no texto extraído, o resultado é um texto padronizado, como visto na figura 16:

Figura 15 - Texto extraído do arquivo pdf com tratamento



Fonte: Dados da pesquisa

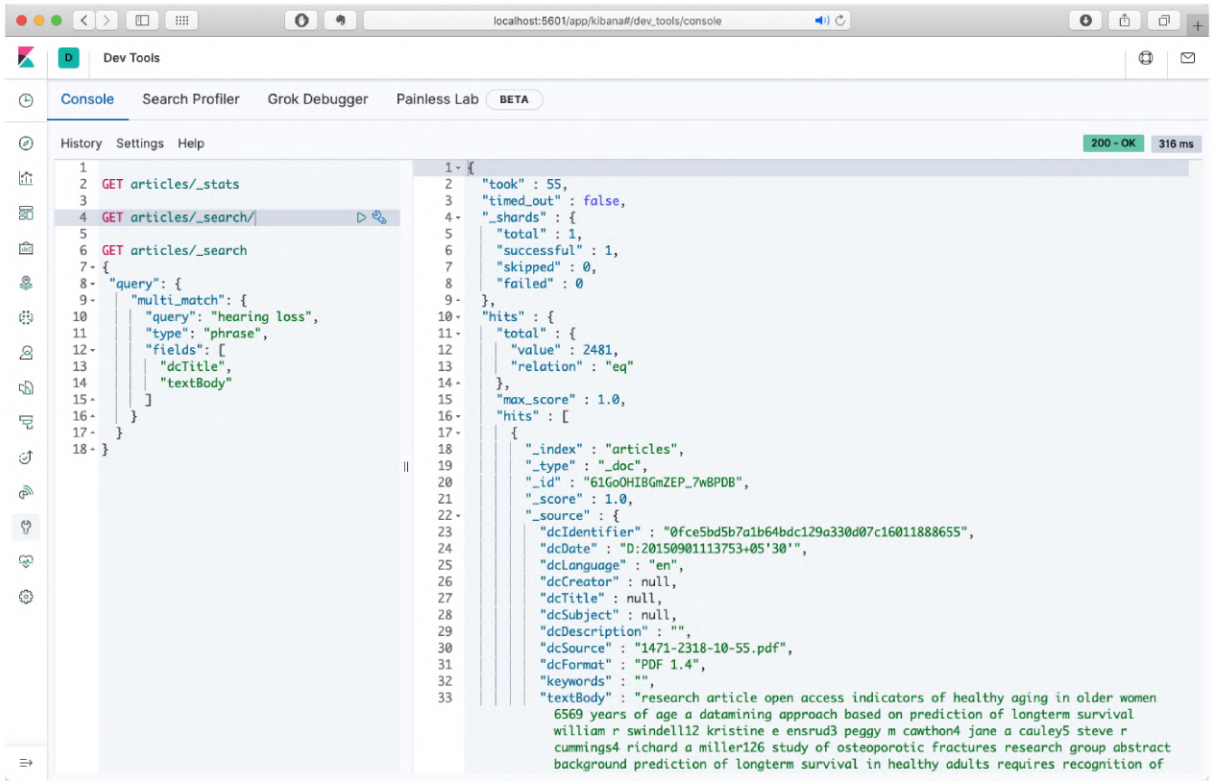
O texto processado apresenta a padronização para caracteres minúsculos (*case-folding*), ausência de quebra de linha, caracteres especiais, pontuação, dentre outras características descritas para o *pipeline* (seção 5.2.2.1). Esse processo é realizado para todos os

arquivos e, a cada interação, o texto é processado juntamente com um conjunto de metadados, depois tudo é armazenado no banco de dados para posterior consulta. Desse processo, portanto, resulta a possibilidade de compatibilização entre a *consultas* pelas terminologias em texto com o texto do artigo.

6.1.1.5 Resultado da estruturação de artigos

A estruturação dos artigos foi uma etapa necessária para viabilizar o acesso aos artigos científicos para que fossem passíveis de pesquisa. A implantação no *Elasticsearch* permitiu o acesso às informações via protocolo REST na porta 5601. A figura 17 ilustra uma interface de alto nível pela ferramenta *Kibana*, parte da suíte de soluções *ElasticSearch*. A interface apoia o desenvolvimento ao enviar comandos, definir estruturas e conferir a recuperação dos artigos. No caso do algoritmo para submissão de *queries*, é feito o uso do banco de dados diretamente, através da biblioteca *Python Elasticsearch Client*[47].

Figura 16 - Fragmento da interface da ferramenta Kibana



Fonte: Dados da pesquisa

[47] https://elasticsearch-py.readthedocs.io/en/master/

O banco de dados em sua definição de esquema foi criado em uma estrutura que permitiu armazenar os artigos científicos e expor uma interface para responder às consultas do algoritmo em Python. A título de demonstração da estatística da estrutura, seguem abaixo trechos, recortados propositalmente, para ilustrar configurações:

Quadro 11 - Estatística estrutural no Elasticsearch

```
{
 "_shards" : {
  "total" : 2,
  "successful" : 1,
  "failed" : 0
 },
...
 "indices" : {
  "articles" : {
   "uuid" : "U4ogVjwJQgSUrRd6nBuQlg",
   "primaries" : {
    "docs" : {
     "count" : 2481,
     "deleted" : 0
    },
    "store" : {
     "size_in_bytes" : 88659572
    },
...
```

Fonte: Dados da pesquisa

Armazenados nesse banco de dados, há 2.481 artigos em texto completo e metadados que totalizaram cerca de 84,5 *MegaBytes* de espaço em disco. Esse artefato, enquanto resultado dos processos de pré-processamento e estruturação de artigos, permite ao algoritmo pesquisar os termos das terminologias e verificar sua revocação através do retorno dos dados em formato JSON.

6.1.2 Resultados do Experimento

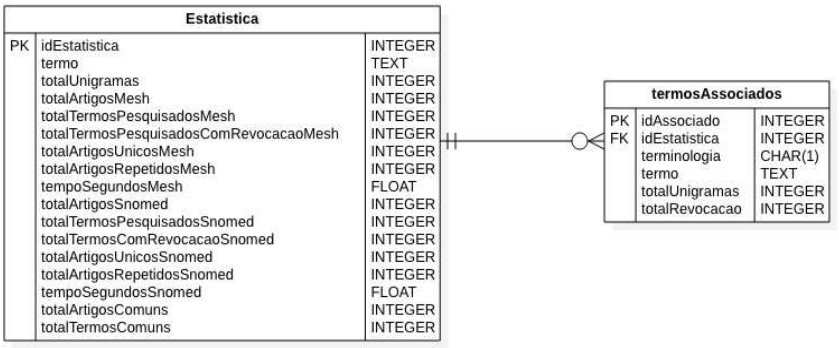
O experimento passou por várias etapas até apresentar os resultados das consultas expandidas conforme ilustrado na figura 4. Apresenta-se nessa seção, o conjunto de etapas, até

o processo final, onde o resultado das consultas permite a avaliação comparativa entre as terminologias estudadas. Os resultados aqui são oriundos de diversas aferições e comparações que vão permitir a discussão e as conclusões.

6.1.2.1 Resultados das consultas expandidas

Quando o *software* realiza a consulta expandida, um grande conjunto de informações é identificado. Esse conjunto, juntamente com a revocação dos artigos, foi armazenado em um banco de dados que definimos como "banco estatístico" (Figura 18). Cada consulta gera um conjunto de dados que são processados e totalizados no *software* desenvolvido, restando ao banco de dados armazenar resultados (Figura 4). Os termos expandidos são armazenados na tabela "termosAssociados", juntamente com sua revocação para cada consulta, a qual, por sua vez recebe um identificador, idEstatistica, que permite o rastreamento dos dados totalizados junto aos termos usados naquela consulta.

Figura 17 - Modelagem do banco de dados para estatística



Fonte: Dados da pesquisa

Assim como ocorreu na estruturação das terminologias, a tecnologia escolhida para esse banco de dados foi o *SQLite*. Abaixo um dicionário de dados a fim de documentar a modelagem (quadro 12).

Quadro 12 - Dicionário de dados para o banco de estatística

Estrutura da Tabela Estatística	
idessatistica	Identificador, chave primária para cada processo de *consulta* por termo comum

termo	Termo comum entre ambas terminologias
totalUnigramas	Número de palavras do termo comum
totalArtigosMesh	Total de arquivos recuperados pelo processo da terminologia MeSH
totalTermosPesquisadosMesh	Total de termos usados no processo da terminologia MeSH
totalTermosPesquisadosComRevocacaoMesh	Dos termos usados no processo MeSH, quantos tiveram revocação
totalArtigosUnicosMesh	Dos artigos recuperados com a MeSH, quantos são únicos (sem repetição)
totalArtigosRepetidosMesh	Dos artigos recuperados com a MeSH, quantos são repetidos
tempoSegundosMesh	Duração em segundos do processamento da consulta expandida
totalArtigosSnomed	Total de arquivos recuperados pelo processo da terminologia SNOMED CT
totalTermosPesquisadosSnomed	Total de termos usados no processo da terminologia SNOMED CT
totalTermoscomRevocacaoSnomed	Dos termos usados no processo SNOMED CT, quantos tiveram revocação
totalArtigosUnicosSnomed	Dos artigos recuperados pela SNOMED, quantos são únicos (sem repetição)
totalArtigosRepetidosSnomed	Dos artigos recuperados com a SNOMED CT, quantos são repetidos
tempoSegundosSnomed	Duração em segundos do processamento da consulta expandida
totalArtigosComuns	Dos arquivos recuperados em ambas terminologias do mesmo processo, quantos artigos são comuns (foram recuperados em ambos processos)
totalTermosComuns	A partir de um mesmo processo de consulta, quantos termos são comuns entre ambas terminologias
Estrutura da Tabela termosAssociados	
idAssociado	Identificador, chave primária para cada registro dessa tabela
idessatistica	Identificador, chave estrangeira para conectar esses dados à tabela estatística
terminologia	“M” para MeSH e “S” para SNOMED CT

termo	Termo expandido (sinônimo, termo de entrada, próximo conceitualmente) a partir do termo comum
totalUnigramas	Quantidade de palavras do termo
totalRevocacao	Total de artigos recuperados por esse termo especificamente

Fonte: Dados da pesquisa

Outro artefato de fundamental importância no projeto foi o *software* desenvolvido em linguagem Python, que possibilitou a execução os processos descritos. Pode-se agrupar o código fonte em cinco grandes grupos, a fim de facilitar o entendimento das principais estruturas do *software*:

Grupo 1- MeSH

- Definição de estrutura BD
- Importação de dados para o BD
- PLN na entrada de dados

Grupo 2- SNOMED CT

- Definição de estrutura BD
- Importação de dados para o BD

Grupo 3- Utilitários

- Download artigos BMC
- Constantes
- Pré-Processamento Textual
- Criação dos Bancos de Dados de Terminologias (*script*)

Grupo 4- *ElasticSearch*

- Popular BD com artigos pdf
- Pesquisa expandida

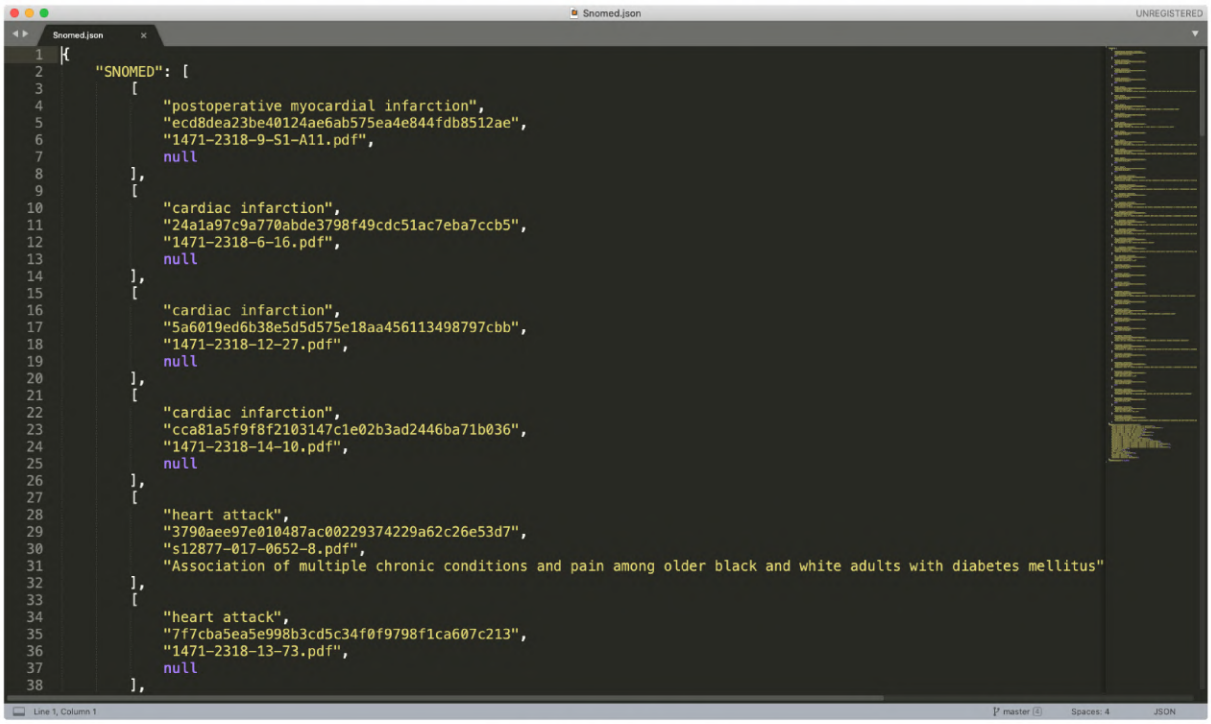
Grupo 4- Flask

- Micro servidor web
- Front end

O *software* foi armazenado no repositório digital *online* GitHub – https://github.com/erfelipe/tese – que deverá ficar disponível para acesso à comunidade com a política de licença *Creative-Commons*, após o devido registro no INPI, órgão nacional de controle de propriedade intelectual. Também estarão disponíveis os *scripts* referentes à criação e estruturação do banco de dados *ElasticSearch* para os artigos científicos.

Demonstra-se abaixo, o procedimento e os dados obtidos a partir de consulta exemplo (vide seção 5.2.2.1). As consultas foram formuladas de modo independente a fim de contemplar a particularidade de cada terminologia. Observa-se que, para cada processo, um conjunto de procedimentos diferentes foi realizado a fim de chegar a um resultado por terminologia, o que, ao final, foi unificado para permitir análise e representação final para o usuário. Apresenta-se abaixo, um exemplo do formato JSON de retorno da pesquisa expandida pelo instrumento SNOMED CT (Figura 19).

Figura 18 - Retorno do processo de query para SNOMED CT, em formato JSON



Fonte: Dados da pesquisa

A partir do conceito de estrutura de dados “chave-valor”, base do formato JSON, o retorno da pesquisa pelas terminologias traz, para cada terminologia, uma lista (Figura 19), onde a chave “SNOMED” possui como valor uma lista de listas. Da mesma forma a consulta com a terminologia MeSH, também retornou uma lista com a chave “MeSH” e como valor, também, uma lista de listas.

Essas listas foram unificadas em um único conjunto, para apresentação na interface, de forma que um código *JavaScript* percorresse os conjuntos e renderizasse uma apresentação em página web. Entretanto, antes de enviar a lista final para o usuário, é feita uma análise para efeito de comparação entre as terminologias. Tal análise se vale de conjunto de dados (quadro 13), onde procurou-se estabelecer parâmetros que vão ajudar na discussão qualitativas sobre as

terminologias. Estes parâmetros foram estabelecidos na medida em que se desenvolvia o *software* e os questionamentos surgiam a respeito do interesse pelos dados.

Quadro 13 - Parâmetros de caráter quantitativo, mas usados para discussão qualitativa

Item de análise estatística	**Descrição**
termo	Termo procurado
totalArtigosMesh	Número de artigos recuperados
totalTermosPesquisadosMesh	Número de termos para consulta expandida
totalTermosPesquisadosComRevocacaoMesh	Número de termos com revocação
totalArtigosUnicosMesh	Número de artigos que não se repetem para os diversos termos pesquisados
totalArtigosRepetidosMesh	Número de artigos que se repetem para os diversos termos pesquisados
totalArtigosSnomed	Número de artigos recuperados
totalTermosPesquisadosSnomed	Número de termos para consulta expandida
totalTermosPesquisadosComRevocacaoSnomed	Número de termos com revocação
totalArtigosUnicosSnomed	Número de artigos que não se repetem para os diversos termos pesquisados
totalArtigosRepetidosSnomed	Número de artigos que se repetem para os diversos termos pesquisados
totalArtigosComuns	Número de artigos que aparecem em revocação para ambas terminologias
totalTermosComuns	Número de termos comuns em ambas terminologias no processo de pesquisa

Fonte: Dados da pesquisa

A fim de alimentar esse conjunto de informações que buscava ser de caráter quantitativo, foram criados dois algoritmos: i) um para identificar, mediante critério, os termos comuns entre as terminologias e armazená-los em lista em arquivo; ii) o outro algoritmo percorre um número pré-determinado de termos e submete-os à pesquisa para gerar um conjunto estatístico para análise. Esta abordagem foi necessária para separar o modelo da

interface de consulta pelo usuário, que possui como característica principal realizar a consulta no *front-end*, desse modelo de objetivos estatísticos em um conjunto de *n* consultas realizadas em *background*.

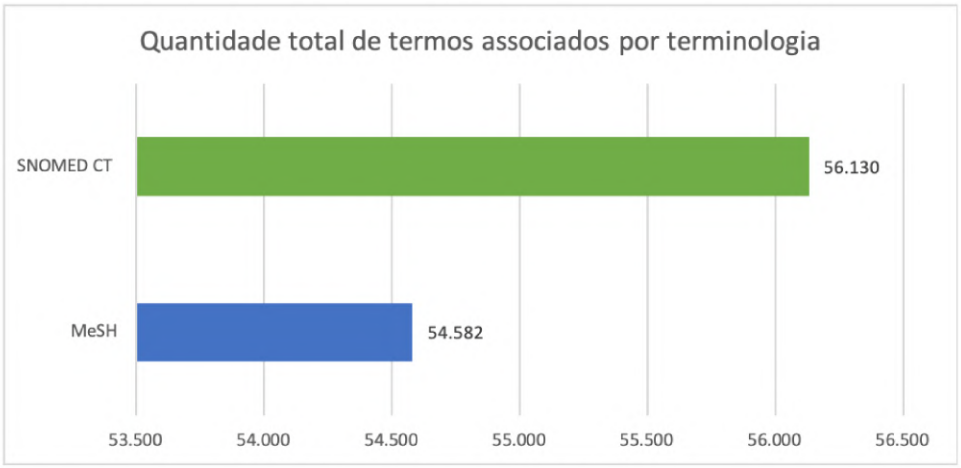
6.1.2.2 Dados estatísticos

Considerando o subconjunto de termos comuns ("*diseases*") em ambas terminologias como princípio norteador das consultas, a análise dos dados comparativos entre a capacidade de revocação dos artigos científicos foi desenvolvida com base em 11 tópicos identificados no restante da presente seção.

A – Quantidade total de termos associados por terminologia

Dos 2.225 termos em comum da categoria "*diseases*", cada terminologia possui a hierarquia de termos, além dos termos próximos conceitualmente (sinônimos), os quais são denominados termos de entrada no MeSH. Esta contagem é relevante por subentender que o número de termos afetou a revocação, mas o levantamento demonstrou certo equilíbrio entre ambas terminologias (Figura 20). Em relação ao quantitativo geral de termos nas terminologias, a SNOMED CT é muito superior. Ainda assim, considerando apenas os termos comuns da categoria citada, a SNOMED CT apresentou aproximadamente mil termos a mais que a MeSH. Considerando a estrutura de conexões hierárquicas, sinônimos, termos de entrada e conceitualmente próximos nas particularidades de cada terminologia, observou-se uma vantagem em termos quantitativos – 1.548 termos – para a SNOMED CT.

Figura 19 - Quantidade de termos amostrais por terminologia

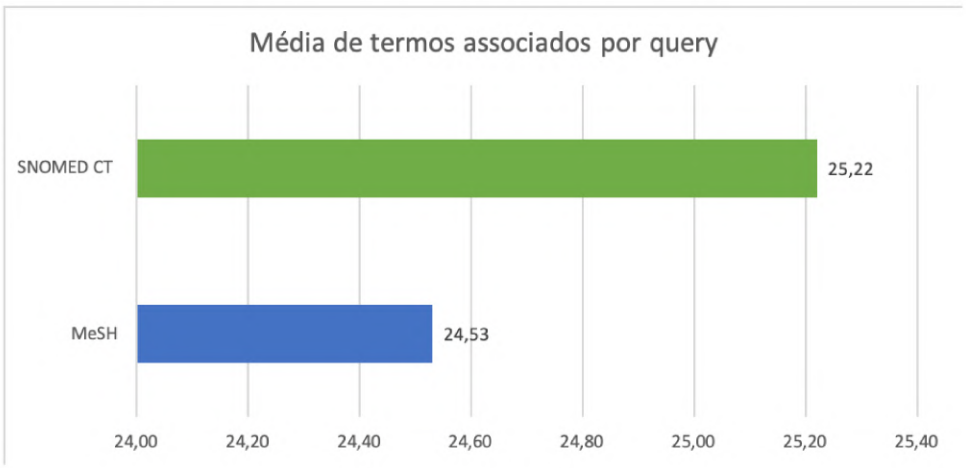


Fonte: Dados da pesquisa.

B – Média de termos associados por consulta

Procurou-se identificar com o conjunto de dados, uma média do conjunto de termos associados à cada consulta, a fim de traçar um parâmetro de distanciamento. A proximidade foi grande, com diferença em média, de apenas aproximadamente um termo por consulta. Quando uma consulta é preparada, ela agrega com base no termo comum, um conjunto de outros termos que são os termos hierárquicos, sinônimos, termos de entrada, etc. Em média, cada consulta de cada terminologia teve um número próximo, com uma leve vantagem quantitativa para a SNOMED CT. Entretanto, a média não demonstra uma distribuição frequente em cada consulta, trata-se apenas um referencial para facilitar o entendimento. Na prática, houve certa discrepância: há termos que permitem uma grande relação de termos associados, de 100, 200, ou mais, enquanto outros com uma dezena ou menos. Dessa forma, destaca-se mais uma vez, a terminologia SNOMED CT no quesito quantidade.

Figura 20 - Média de termos associados por consulta

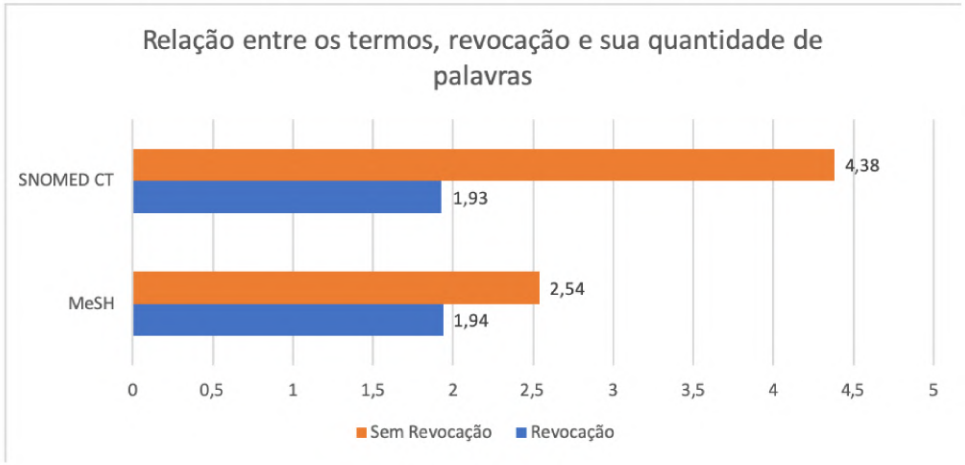


Fonte: Dados da pesquisa.

C – Relação entre os termos, revocação e quantidade de palavras

Houve uma nítida relação entre termos que obtiveram revocação e o quantitativo de palavras presentes. Pôde-se inferir claramente que quanto maior o número de palavras, mais deficiente foi a revocação. Os dados apontam que as revocações ocorreram, em média, com termos de duas palavras.

Figura 21 - Relação entre termos e revocação

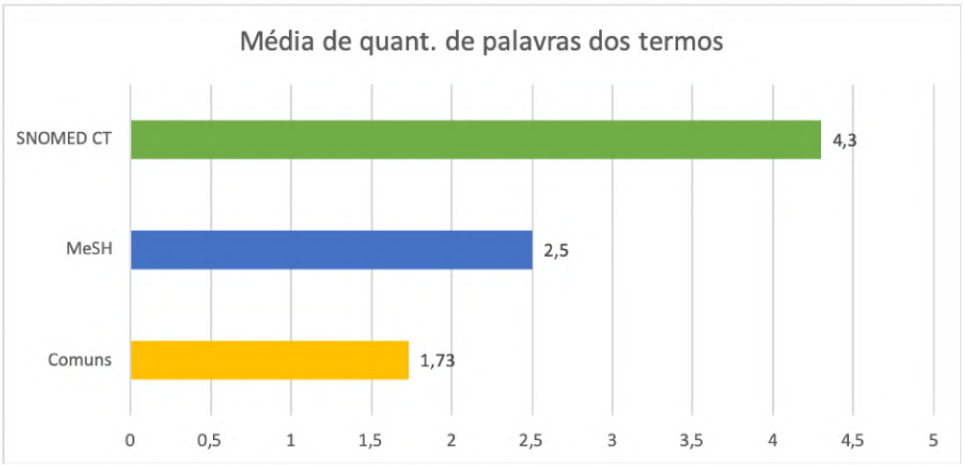


Fonte: Dados da pesquisa.

D – Média de quantidade de palavras nos termos por terminologia

Uma análise segmentada demonstra de forma mais clara que as terminologias possuem características distintas em sua construção terminológica. Enquanto a SNOMED CT utiliza termos com um número maior de palavras, a MeSH indica a construção com um menor número de palavras por termo. A Figura 23 indica também os termos comuns entre ambas terminologias, que possuem em média duas palavras.

Figura 22 – Média de quantidade de palavras nos termos

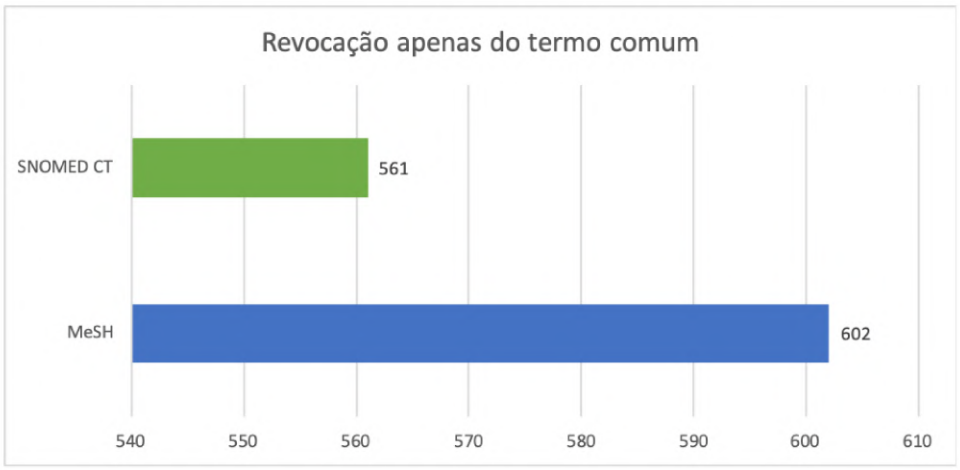


Fonte: Dados da pesquisa.

E – Número de consultas que possuem revocação apenas para o termo comum

Nesta análise considerou-se apenas revocações do termo comum entre as terminologias quando os termos associados não retornaram revocação. Ou seja, o termo comum foi o responsável único pela revocação, tornando-se a representação capaz de recuperar os artigos mesmo quando os demais termos associados não o fizeram.

Figura 23 - Revocações apenas pelo termo comum

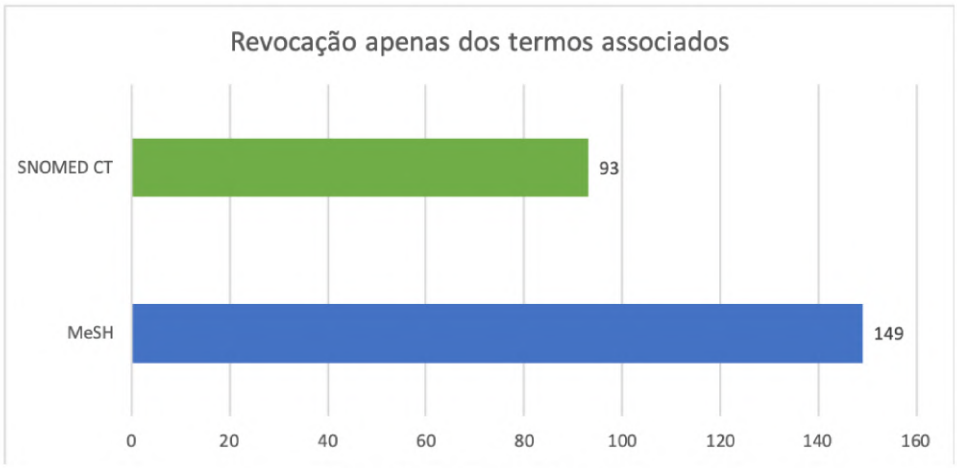


Fonte: Dados da pesquisa.

F – Número de consultas que possuem revocação apenas para os termos associados

Esta análise é oposta à anterior. No gráfico pôde-se observar os momentos onde o termo comum não teve revocação, a qual foi gerada por algum termo associado, por grupo de consulta.

Figura 24 - Revocação apenas pelos termos associados

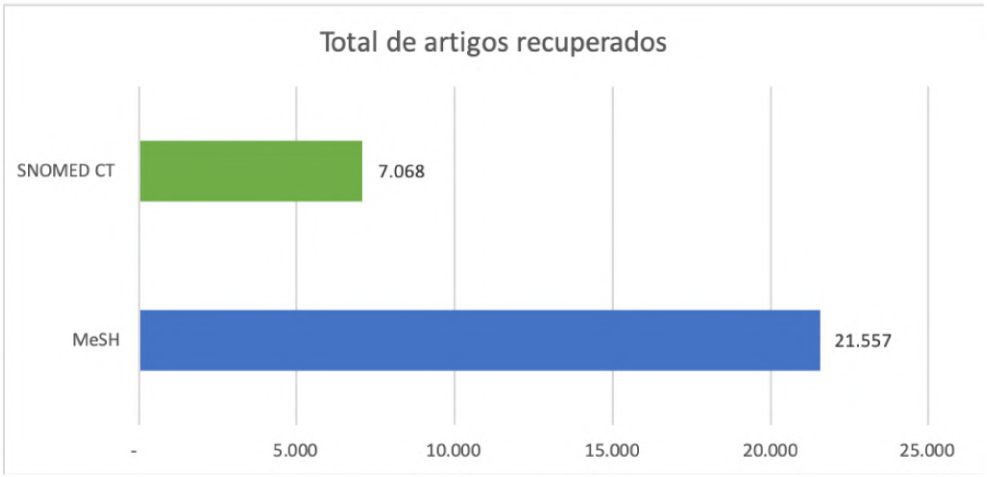


Fonte: Dados da pesquisa.

G – Total geral de artigos recuperados

Em um somatório geral, considerando a quantidade de artigos recuperados por terminologia, incluindo artigos repetidos no mesmo conjunto da consulta, os números apontam discrepância: a MeSH obtém uma dimensão três vezes maior em revocação, no quantitativo de artigos.

Figura 25 - Total de artigos recuperados

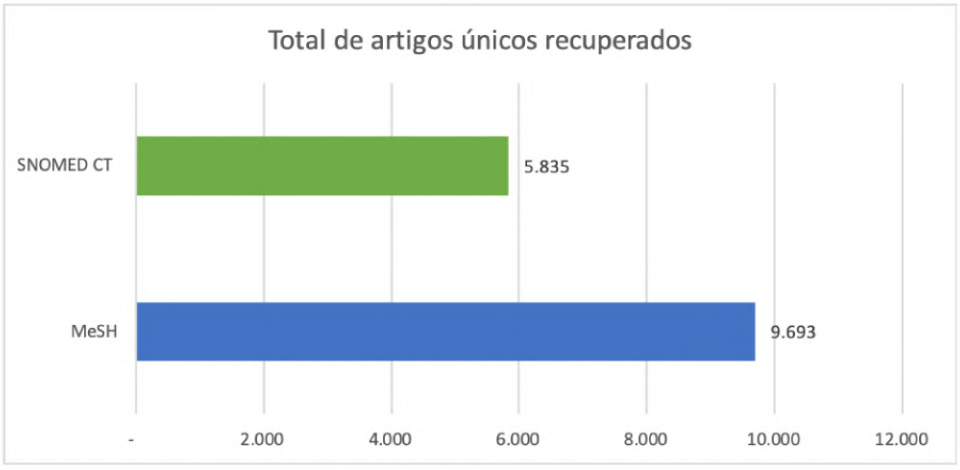


Fonte: Dados da pesquisa.

H – Total geral de artigos únicos recuperados

A análise permite observar que, mesmo com uma totalização de revocação em arquivos únicos (sem repetição por consulta), uma das terminologias se destaca em quantitativo.

Figura 26 - Total de artigos únicos

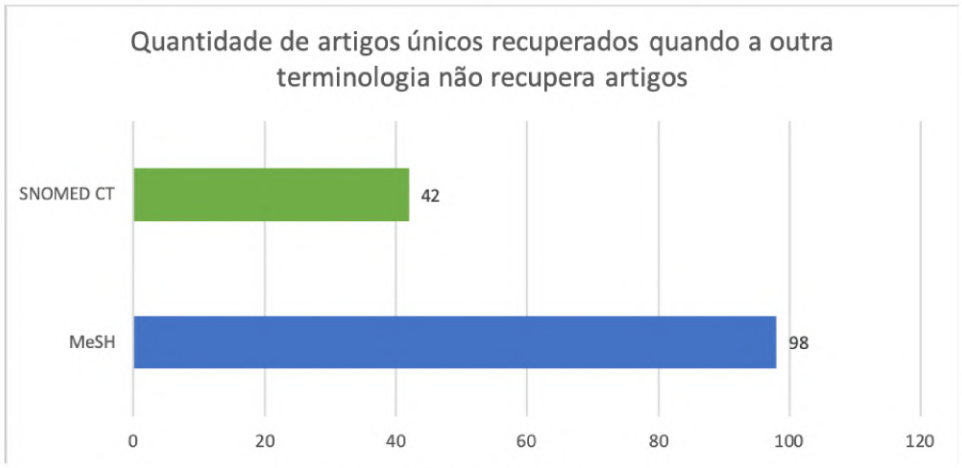


Fonte: Dados da pesquisa.

I – Artigos únicos recuperados quando a outra terminologia não recupera

Esse levantamento demonstra a situação onde uma terminologia recupera artigos, com termo comum ou associado, e a outra terminologia não recupera, para o mesmo grupo a partir do termo comum.

Figura 27 - Quantidade de *queries* que retornaram artigos únicos quando a outra terminologia não retornou resultado

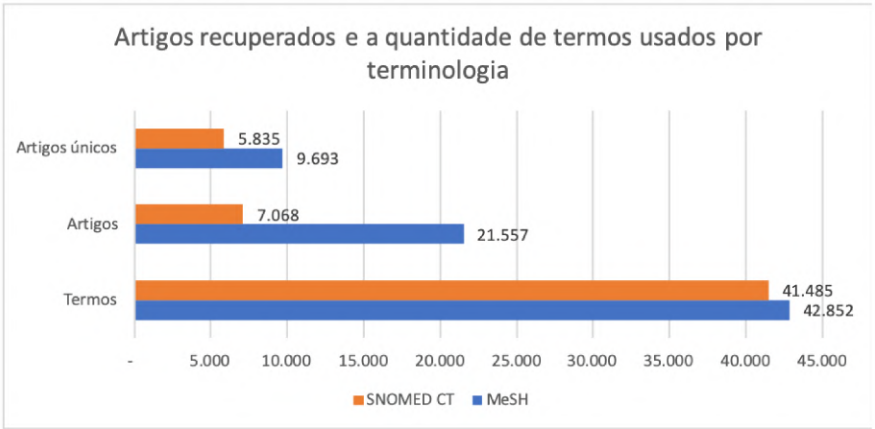


Fonte: Dados da pesquisa.

J – Artigos recuperados e quantidade de termos com revocação

Esse levantamento é um compilado com a adição da quantidade de termos. Esse somatório tem como base a presença de revocação, visto que sem o critério de revocação, a SNOMED CT possui mais termos. A quantidade de termos usados e a quantidade de artigos recuperados pode nortear uma análise de proximidade na melhor representação a fim de recuperar os documentos científicos.

Figura 28 - Artigos e quantidade de termos

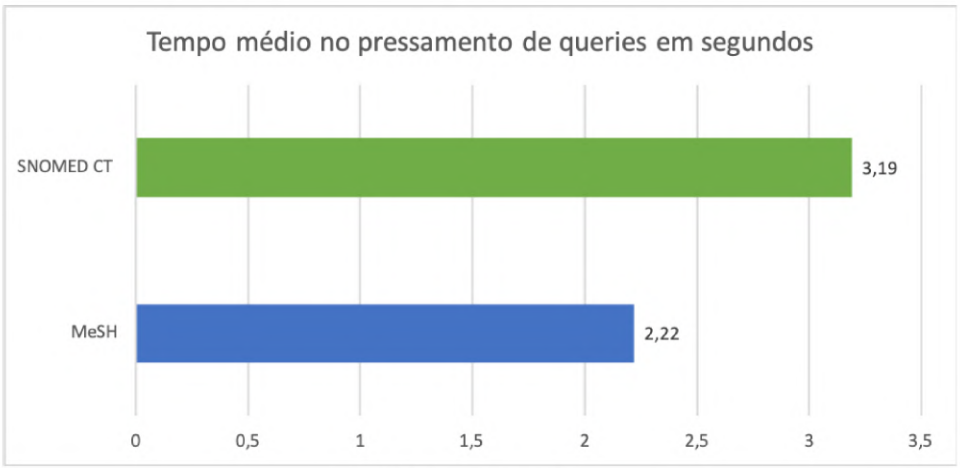


Fonte: Dados da pesquisa.

K – Tempo médio para o processamento das consultas, em segundos

Esta análise permite observar claramente que o tempo médio para processamento das consultas pela SNOMED CT é, em média, um segundo mais lento que as consultas realizadas na MeSH.

Figura 29 - Tempo médio de submissão da query



Fonte: Dados da pesquisa.

6.2 Discussão

O desenvolvimento do trabalho de pesquisa em geral encontra limitações. É hora de elencar itens que impactaram o desenvolvimento da pesquisa. Esses itens são discutidos ao longo da presente seção em: i) limitações (Seções 5.2.1); ii) problemas e desvios (Seções 5.2.2); iii); terminologias vs RI (seção 5.2.3). Finalmente, apresenta-se uma discussão global sobre os resultados da pesquisa (seção 5.2.4).

6.2.1 Limitações

Uma limitação da pesquisa envolve extração do texto dos artigos. Conforme descrito na Seção 4.2.2, uma preocupação do projeto recai sobre a viabilidade de criar o banco de dados de artigos, em formato de texto completo. Apesar de várias bibliotecas de extração terem sido testadas, nenhuma apresentou extração livre de erros ou de situação que não demandasse algum nível de tratamento.

Foi escolhida uma biblioteca de extração, avaliada a mais adequada ao projeto conforme testes já descritos. O corpo textual, portanto, não figura em um estado perfeitamente original, mas sim em estado viável para o processo de pesquisa texto completo. O ideal seria ter tido acesso ao documento antes de sua produção em pdf, o que não era viável. Isso exigiu processos para o tratamento do material a fim de retirar quebras de linhas, hifenização, caracteres especiais, pontuação, dentre outros. Só assim foi possível implementar o banco de dados passível de responder às consultas.

Outro tipo de limitação envolveu o pré-processamento. Por ser um processo crucial para a viabilidade do projeto, os problemas no pré-processamento dos artigos eram impactantes. Ao tratá-los, foi possível obter um texto mais próximo do original com possibilidade de pesquisa texto completo. Como já mencionado, o formato pdf objetiva a leitura humana, mas é inadequado para a RI automatizadas. Para pesquisar nesse formato era necessário um sistema de arquivos ou um banco de dados que armazenaria não o texto, mas o arquivo pdf. São abordagens tecnicamente passíveis de implementação, mas de alto custo computacional, no sentido do tempo de processamento. Adiciona-se a esse cenário, o fato que seria necessário o acesso via uma API para pesquisa em arquivos pdf, além do processo de abertura de arquivo, o que tornaria a consulta em tempo real inviável.

Desse modo, a estratégia adotada no presente trabalho, de extração e tratamento do conteúdo textual em texto, mostrou-se satisfatória para o processo de consulta expandida.

6.2.2 Problemas e desvios

A presente seção destaca as dificuldades encontradas no decorrer do desenvolvimento do *software*, bem como as soluções. Envolve as etapas de aquisição de terminologia e de extração de texto dos artigos.

Com relação a aquisição de uma terminologia, faz-se necessário esclarecer que o processo de aprovação na instituição que mantém a terminologia SNOMED CT foi muito lento, em torno de 60 dias entre o preenchimento do formulário e a autorização. Isso impactou o projeto e gerou insegurança sobre a possibilidade de uso da terminologia no projeto.

Com relação a extração do texto dos artigos, um problema afetava diretamente a qualidade de RI era a extração do texto em pdf para compor o banco de dados. Mesmo que a primeira seleção da biblioteca de extração tenha sido considerada a melhor solução, mantinha-se um comportamento indesejado: de forma aleatória, algumas palavras apareceriam sem

espaços (junção de palavras). De forma a não impactar no andamento dos outros processos do *software*, os demais algoritmos foram criados mesmo com a base de dados com essa fragilidade.

Porém, ao revisitar o processo, descobriu-se uma nova biblioteca que apresentou extração aceitável, sem a ocorrência de junção de palavras. Os algoritmos de exclusão de outros casos – hifens, URLs, endereços de e-mail, pontuações e termos especiais – permaneceram inalterados. A adoção da nova biblioteca exigiu uma modificação mínima do algoritmo na parte que trata da leitura dos metadados, visto que o dicionário recuperado de ambas bibliotecas possui particularidades.

Como a base de dados foi construída com a pretensão de detalhamento além do corpo textual, de forma a permitir maior especificidade na RI no futuro, alguns campos foram incluídos para alcançar tal granularidade. Esses campos estão detalhados na seção 5.2.2.1, mas aqui destacam-se as possibilidades de identificação do título e do autor pela leitura dos metadados nos arquivos pdf. A título de exemplo, apresenta-se abaixo a extração das bibliotecas para o mesmo arquivo, a começar pela biblioteca *pyMuPdf*:

Quadro 14 - Metadados pyMuPdf

{“format”: “pdf 1.4”, “**title**”: “**Improvement of pressure ulcer prevention care in private for-profit residential care homes: an action research study**”, “**author**”: “Enid WY Kwong”, “subject”: “BMC Geriatrics, 2016, doi:10.1186/s12877-016-0361-8”, “keywords”: “Action research,Pressure ulcer prevention,Residential care homes,Gerontology”, “creator”: “Arbortext Advanced Print Publisher 9.1.440/W Unicode”, “producer”: “Acrobat Distiller 10.1.5 (Windows); modified using iText® 5.3.5 ©2000-2012 1T3XT BVBA (AGPL-version)”, “creationDate”: "D:20161124023041+08”00”", “modDate”: "D:20161125020652+01”00”", “encryption”: None}

Fonte: Dados da pesquisa

Já os metadados do mesmo arquivo pela biblioteca *Apache Tika-python* são:

Quadro 15 - Metadados Tika-Python

{“Author”: [“Enid WY Kwong”, “Maria SY Hung”, “Kevin Woo”], “Content-Type”: “application/pdf”, “Creation-Date”: “2016-11-23T18:30:41Z”, “CrossMarkDomains[1]”: “springer.com”, “CrossMarkDomains[2]”: “springerlink.com”, “CrossmarkDomainExclusive”: “true”, “CrossmarkMajorVersionDate”: “2016-11-24”, “Keywords”: “Action research,Pressure ulcer prevention,Residential care homes,Gerontology”, “Last-Modified”: “2016-11-25T01:06:52Z”, “Last-Save-Date”: “2016-11-25T01:06:52Z”, “X-Parsed-By”: [“org.apache.tika.parser.DefaultParser”, “org.apache.tika.parser.pdf.PDFParser”], “X-TIKA:content_handler”: “ToTextContentHandler”, “X-TIKA:embedded_depth”: “0”, “X-TIKA:parse_time_millis”: “181”, “access_permission:assemble_document”: “true”, “access_permission:can_modify”: “true”, “access_permission:can_print”: “true”, “access_permission:can_print_degraded”: “true”, “access_permission:extract_content”: “true”, “access_permission:extract_for_accessibility”: “true”, “access_permission:fill_in_form”: “true”, “access_permission:modify_annotations”: “true”, “cp:subject”: “BMC Geriatrics, 2016, doi:10.1186/s12877-016-0361-8”, “created”: “2016-11-23T18:30:41Z”, “creator”: [“Enid WY Kwong”, “Maria SY Hung”, “Kevin Woo”], “date”: “2016-11-25T01:06:52Z”, “dc:creator”: [“Enid WY Kwong”, “Maria SY Hung”, “Kevin Woo”], “dc:description”: “BMC Geriatrics, 2016, doi:10.1186/s12877-016-0361-8”, “dc:format”: [“application/pdf; version="A-2b"“, “application/pdf; version=1.4”], “dc:language”: “EN”, “dc:subject”: “Action research,Pressure ulcer prevention,Residential care homes,Gerontology”, “**dc:title**”: “**Improvement of pressure ulcer prevention care in private for-profit residential care homes: an action research study**”, “dcterms:created”: “2016-11-23T18:30:41Z”, “dcterms:modified”: “2016-11-25T01:06:52Z”, “description”: “BMC Geriatrics, 2016, doi:10.1186/s12877-016-0361-8”, “doi”: “10.1186/s12877-016-0361-8”, “language”: “EN”, “**meta:author**”: [“Enid WY Kwong”, “Maria SY Hung”, “Kevin Woo”], “meta:creation-date”: “2016-11-23T18:30:41Z”, “meta:keyword”: “Action research,Pressure ulcer prevention,Residential care homes,Gerontology”, “meta:save-date”: “2016-11-25T01:06:52Z”, “modified”: “2016-11-25T01:06:52Z”, “pdf:PDFVersion”: “1.4”, “pdf:charsPerPage”: [“3041”, “4711”, “4613”, “3353”, “4656”, “4533”, “4794”, “4629”, “4833”, “4997”, “4563”, “4598”, “6123”, “4487”], “pdf:docinfo:created”: “2016-11-23T18:30:41Z”, “pdf:docinfo:creator”: “Enid WY Kwong”, “pdf:docinfo:creator_tool”: “Arbortext Advanced Print Publisher 9.1.440/W Unicode”, “pdf:docinfo:custom:CrossMarkDomains[1]”: “springer.com”, “pdf:docinfo:custom:CrossMarkDomains[2]”: “springerlink.com”, “pdf:docinfo:custom:CrossmarkDomainExclusive”: “true”, “pdf:docinfo:custom:CrossmarkMajorVersionDate”: “2016-11-24”, “pdf:docinfo:custom:doi”: “10.1186/s12877-016-0361-8”, “pdf:docinfo:keywords”: “Action research,Pressure ulcer prevention,Residential care homes,Gerontology”, “pdf:docinfo:modified”: “2016-11-25T01:06:52Z”, “pdf:docinfo:producer”: “Acrobat Distiller 10.1.5 (Windows); modified using iText® 5.3.5 ©2000-2012 1T3XT BVBA (AGPL-version)”, “pdf:docinfo:subject”: “BMC Geriatrics, 2016, doi:10.1186/s12877-016-0361-8”, “pdf:docinfo:title”: “Improvement of pressure ulcer prevention care in private for-profit residential care homes: an action research study”, “pdf:encrypted”: “false”, “pdf:hasMarkedContent”: “false”, “pdf:hasXFA”: “false”, “pdf:hasXMP”: “true”, “pdf:unmappedUnicodeCharsPerPage”: [“0”, “0”, “0”, “0”, “0”, “0”, “0”, “0”, “0”, “0”, “0”, “0”, “0”, “0”, “0”], “pdfa:PDFVersion”: “A-2b”, “pdfaid:conformance”: “B”, “pdfaid:part”: “2”, “producer”: “Acrobat Distiller 10.1.5 (Windows); modified using iText® 5.3.5 ©2000-2012 1T3XT BVBA (AGPL-version)”, “resourceName”: "b”s12877-016-0361-8.pdf”", “subject”: “BMC Geriatrics, 2016, doi:10.1186/s12877-016-0361-8”, “title”: “Improvement of pressure ulcer prevention care in private for-profit residential care homes: an action research study”, “xmp:CreatorTool”: “Arbortext Advanced Print Publisher 9.1.440/W Unicode”, “xmpMM:DocumentID”: “uuid:66aa5abf-1548-4e41-bfc9-6245cff24cb6”, “xmpMM:History:Action”: “converted”, “xmpMM:History:InstanceID”: “uuid:4e034033-f520-49c0-84dc-2d8227e3075a”, “xmpMM:History:SoftwareAgent”: “pdfToolbox”, “xmpMM:History:When”: “2016-11-24T02:38:10Z”, “xmpTPg:NPages”: “14”}

6.2.3 Terminologias vs RI

Os artefatos terminológicos escolhidos MeSH e SNOMED CT são terminologias clínicas respeitadas mundialmente. As escolhas seguiram critérios (vide Seção 2.2) e, embora sejam terminologias da mesma área de conhecimento, suas características (Seções 2.2.1 e 2.2.2) permitiram uma análise diferenciada no processo de RI. Descrevem-se aqui 3 aspectos

relacionados a terminologias: i) o investimento em terminologias; ii) o processamento em tempo real; e iii) características das terminologias.

Um aspecto importante é a discussão sobre a continuidade dos investimentos em terminologias. Em 1972, Salton já tinha publicado trabalho de comparação entre um modelo tradicional de indexação (MEDLARS) e o modelo de processamento textual automático (SMART). Afirmava que, naquele momento, a adição de artefatos como uma lista controlada ou um tesauro no modelo automático (SMART) tornava resultados de RI comparáveis ao modelo intelectual (MEDLARS) e que o retorno de usuários incrementou o modelo automático em percentuais consideráveis. Mais tarde, o modelo vetorial transformaria o processo de RI e criaria um novo paradigma no processamento automático. Em 1996, Salton indicava extrema confiança ao afirmar que o modelo estatístico sobrepujava outros modelos, indicando que o horizonte da CI precisava se ampliar nesse rumo, a ponto de rotular artefatos como vocabulários controlados e tesauros, como misérias do século XIX (SALTON, 1996).

Entretanto, cabe refletir se seria realmente o caso de desconsiderar todo o esforço despendido na criação e manutenção desses artefatos terminológicos. O presente trabalho leva a crer que, realmente, o modelo estatístico suplantou os demais modelos, mas em situações de domínio específico, em áreas dependentes de terminologias complexas, ainda há lugar para que os artefatos terminológicos possam contribuir de modo significativo no processo de RI.

A própria ISKO-UK em 2015 propôs um debate baseado na afirmativa “esta casa acredita que não há mais lugar para o tesauro tradicional nos modernos sistemas de recuperação” (CLARKE, 2106). Realmente, o sistema estatístico na proposição abrangente de RI é o paradigma que vingou, mas tesauros, ontologias, ou seja, terminologias de domínio, seriam agora dispensáveis?

Com relação ao processamento em tempo real, o uso de terminologias em sistemas computacionais deve levar em consideração o tempo de processamento para:

- Acessar a terminologia (carregamento em memória)
- Identificação de termos hierárquicos
- Identificação de termos de entrada e/ou próximos
- Submissão do conjunto de termos ao banco de dados
- Retornar os resultados

Esse tempo, adicionado a outros processos relacionados a interface, uma vez dilatados, podem inviabilizar o uso das terminologias com a expansão de consultas. No experimento aqui produzido, realizou-se juntamente ao processamento de consultas uma medição para identificar o tempo do processamento de cada terminologia em relação à mesma

consulta. Os dados apontaram a seguinte média: MeSH, 2.2 segundos e SNOMED, 3.19 segundos. Esse resultado evidencia que uma terminologia chega a ser quase 1 segundo mais "cara" que outra em termos computacionais. A medição está associada a outros fatores, característicos de cada terminologia, como quantitativo de palavras por termo, expressividade e relações hierárquicas, etc. A análise é apresentada adiante, na discussão sobre os resultados da pesquisa.

Finalmente, cabe citar algumas características das terminologias clínicas que foram observadas. A MeSH, por exemplo, deixa evidente uma falta de padrão na construção terminológica:

- Termos com caracteres especiais
- Termos com operadores booleanos
- Termos com diferentes formatos (maiúsculas ou minúsculas)

Ao verificar o histórico dessa terminologia, observou-se a ausência de critérios rigorosos quando a outros artefatos que foram agregados no passado. A política que permitiu a aquisição e fusão de outras terminologias com o MeSH, teoricamente, deveria primar por modo criterioso, e ser baseada por especialistas na análise de termos, relações e forma sintática. Acredita-se que tal validação seria um fator importante para minimizar os impactos advindos da absorção de novos termos sem devido tratamento para padronizar o conjunto terminológico.

A MeSH foi criada a partir de terminologias que tinham como objetivo a representação documental, o que sem dúvida traçou um perfil favorável à RI. Entre as primeiras terminologias que fizeram a composição da MeSH estavam listas de cabeçalhos de assunto e catálogos. Essa composição indica um importante suporte para representação, catalogação e indexação de documentos médicos.

Embora esse método de construção da terminologia possa ter minimizado o tempo de desenvolvimento e trazido a experiência de trabalhos de equipes diferentes, incorporou características que diversificaram os termos em sua forma de apresentação, nem sempre de forma adequada. Essas características não favorecem o processo de RI, pois são diversas à concepção autoral de documentos científicos. Como exemplo de termos diferenciados, pode-se citar:

- Inversão sintática: "*abnormalities, multiple*" em contraponto a termos sem inversão como "*congenital abnormalities*";
- Múltipla significação: "*diet, food, and nutrition*" e "*human papillomavirus recombinant vaccine quadrivalent, types 6, 11, 16, 18*", "*dna-(apurinic or apyrimidinic site) lyase*" dentre outros;

- Codificação junto ao termo: "*o-(chloroacetylcarbamoyl)fumagillol*";
- Termos não médicos: "*africa, southern*", "*electronic data processing*";
- Termos veterinários: "*african horse sickness*", "*animals*", "*animal identification systems*";
- Operadores booleanos junto ao termo: "*oxidoreductases acting on aldehyde or oxo group donors*", "*heterocyclic compounds, 4 or more rings*", "*oil and gas industry*".

Já a terminologia SNOMED CT apresentou características distintas da MeSH, mesmo sendo também uma terminologia da área médica. Pode-se considerar que a SNOMED CT é uma terminologia "refinada" no sentido em que sua construção está além da organização hierárquica tradicional. Possui três componentes: conceitos, descrições e relacionamentos. Não se trata apenas de terminologia para diagnósticos, mas cobre todo um conjunto de descobertas clínicas, perpassando diagnósticos, sintomas, procedimentos, observações, estrutura corporal, dentre outros. Tal característica ficou nítida na execução do algoritmo do experimento.

A SNOMED é bem mais extensa do que a MeSH. Entretanto, enquanto os termos da MeSH possuem em média três palavras (trigramas), os termos da SNOMED CT possuem cinco palavras. Ou seja, trata-se de uma terminologia também mais "verbosa", que se aproxima mais dos atos médicos do que a representação para classificação e indexação da MeSH. E assim como a MeSH, a SNOMED CT também traz problemas nas descrições que dificultam a RI:

- Inversão sintática: "*application of dressing, major*";
- Múltipla significação: "*(autoantibody titres nos) or (anti-thrombin iii level)*";
- Caracteres especiais: "*% daily total food energy intake*", "*#leg - tibia/fibula*", "*#ankle (& [tibia])*", "*(+)-sabinol dehydrogenase*", "*forces recruit medical (& [admin] or [mod f/med/1])*";
- Expressões booleanas: "*((marital: [conflict] or [disharmony]) (& [row with wife]))*".

Foi possível perceber, no entanto, que a quantidade de termos da terminologia, por ser maior que a MeSH, diminuiu os efeitos negativos no processo de RI. O quantitativo de termos nas versões usadas no desenvolvimento desse trabalho no ano de 2020 foram mensuradas pela consulta: "select count (<campo>) from <tabela>" da seguinte forma:

- Para o SNOMED CT foram totalizados 352.568 conceitos e 1.216.589 descrições;
- No MeSH (sem adição de suplementos), foram totalizados 29.640 descritores principais e 150.349 termos de entrada, que somados formam 179.989 termos.

O MeSH representa, cerca de 15% da quantidade total de termos da SNOMED CT. Salienta-se, portanto, que embora o quantitativo total demonstre uma grande diferença, os termos comuns entre ambas terminologias foram filtrados em um subconjunto a partir da

categoria "*Disease*". Essa identificação selecionou 2.225 termos de sintaxe comum. A partir destes termos foram considerados 56.130 termos associados na terminologia SNOMED e 54.582 na terminologia MeSH.

6.2.4 Discussão dos resultados da pesquisa

A discussão sobre os resultados de pesquisa aborda cinco aspectos principais, à saber: i) as terminologias; ii) a infraestrutura de artigos; iii) a estruturação do banco de dados para estatística; iv) o pré-processamento; e v) resultados da estatística.

Com relação às terminologias, a estruturação em um banco de dados no modelo de entidade-relacionamento resultou em aspectos relevantes no desenvolvimento do experimento de *software*. Destacam-se, dentre outras, as seguintes características:

- Padronização das consultas na identificação de termos
- Velocidade de recuperação dos termos por índices internos
- Velocidade na recuperação de estruturas hierárquicas
- Possibilidades de tratamento textual (não usado nesse experimento)
- Possibilidade de recuperação *cross / multilingual* em vários idiomas

Embora pudessem ser unificadas as informações das terminologias em um mesmo banco de dados, com suas tabelas e índices separados logicamente, cada terminologia foi estruturada em um banco independente. Isso facilitou a manutenção do *software*, visto que as terminologias passaram por atualizações periódicas, além de testes locais constantes. De forma que a estrutura era redefinida, a cada evento, sem impactar em outra ou nas demais partes do sistema.

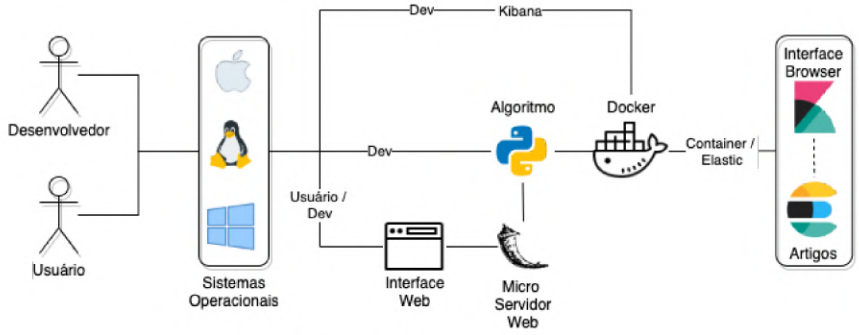
Acrescenta-se também a escolha de banco de dados SQL independente de *software* servidor de banco de dados. O *SQLite* foi ideal para esse tipo de implementação pois permitiu consultas em SQL com conexão direta pela biblioteca Python *sqlite3*[48]. Entende-se que esse modelo atendeu, enquanto projeto de pequena escala, mas em projetos maiores deve-se analisar a perspectiva de utilização de bancos de dados mais robustos.

Com relação a infraestrutura dos artigos, utilizou-se abordagem diferente em comparação com as terminologias. A opção foi usar um banco de dados que faz parte de um conjunto de soluções em RI. O *Elasticsearch*, em sua camada de banco de dados, é um modelo orientado a documentos como alternativa ao paradigma relacional. Isso se traduz em um banco

[48] Disponível na Internet em: https://docs.python.org/3/library/sqlite3.html

que não requer estruturas de dados rígidas e pré-definidas, características que norteia outros bancos de dados no modelo entidade relacionamento. Uma visão gráfica e estrutural do modelo adotado no experimento pode ser vista na figura 42 abaixo:

Figura 42 - Modelo estrutural de acesso aos artigos



Fonte: Dados da pesquisa

Nesse modelo, são possíveis dois tipos de acesso ao banco de dados: i) o acesso pelo desenvolvedor de *software* (dev); ii) pelo usuário final. Destaca-se as características de ambos os acessos:

i) O desenvolvedor pode manipular o *Elasticsearch* por duas vias: pela linguagem de programação, ou pela interface em browser, exposta pela ferramenta *Kibana* na porta 5601, via *Browser*. Na primeira opção, a linguagem de programação escolhida deve acesso de forma nativa ou fazendo o uso de bibliotecas externas. Nesse trabalho, fez-se uso de uma biblioteca específica para *Python*, desenvolvida pela própria *Elasticsearch*. O processamento para fins estatísticos foi feito dessa forma. Na segunda opção, o *Elasticsearch* é usado via camada de *software Kibana* acessada por browser. Esse tipo de interface é importante para questões de criação do índice, mapeamento inicial do modelo de documentos (*schema*), inserção de dados de forma manual, testes e monitoramento.

ii) Ao usuário final foi designado o acesso somente por *Browser*, em interface de alto nível. Utilizando um micro servidor web (Framework *Flask*), ofereceu-se uma interface HTML que dispara uma requisição para a porta 9200 via http. O *Elasticsearch* então é capaz de receber a consulta, processar os dados e retornar um objeto JSON para o browser, cuja página é responsável por apresentar as informações ao usuário por *JavaScript*. O modelo de acesso pelo usuário final é uma estrutura a ser usada para demonstração e futuras implementações do projeto em modelos reais.

Com relação a estruturação do banco de dados para estatística, o experimento se desenvolveu na comparação dos processos de revocação em cada terminologia. Os resultados foram armazenados em uma estrutura de banco de dados com a mesma tecnologia usada nos bancos de dados de terminologias (SQLite). Uma nova estrutura foi criada também com base no modelo entidade-relacionamento. Dessa forma, os dados para análise ficaram disponíveis nessa estrutura de dados, onde a tabela "estatística" possui os principais dados para avaliação, com alguns campos totalizados para facilitar o entendimento. A tabela "termosAssociados" permite conferir aos termos hierárquicos, a proximidade conceitual e a revocação por termo, de cada terminologia. Em todos os termos há uma contagem de palavras – unigramas, bigramas, trigramas, etc. – a fim de permitir inferências de acordo com as características dos resultados.

No que diz respeito ao pré-processamento dos artigos cabe destacar alguns desafios. Foram testadas diversas bibliotecas de *software* para realizar a extração de texto puro do formato proprietário pdf. Para ter uma definição de qual serviço de *software* seria o mais adequado, foi construído um projeto à parte. Nesta iniciativa foram instaladas e testadas as seguintes bibliotecas: *pymupdf*, *pdftotext*, *tika-python* e *pypdf2*. Tendo em vista que o principal critério para escolha era a fidelidade textual entre o texto puro resultante e o artigo em formato pdf.

Ao considerar a formatação dos artigos obtidos em sua apresentação em colunas, percebeu-se que tal diagramação é útil para a leitura humana, mas desnecessária e mesmo inconveniente para o processamento computacional. Nesse contexto, mesmo algoritmos mais eficientes apresentaram deficiências como: reconhecimento errado de caracteres, troca de ordem textual, principalmente em textos com formatação em colunas, além de inserção de símbolos não pertinentes à redação original. Acrescenta-se o fato de que a extração não reconhecia a grafia da divisão silábica por hífen, e considerava uma palavra como duas termos independentes. Esses problemas foram potencializados em documentos que continham tabelas, marcação de tópicos enumerados ou não, e imagens acompanhadas de rótulos e legendas. Esses elementos são comuns em formatação de artigos científicos e conferem complexidade a conversão textual.

O sucesso do processo configurava-se em uma situação crítica para a continuidade do projeto, pois a busca textual é baseada na comparação da sintaxe terminológica. Nas bibliotecas de extração utilizadas, a organização do resultado final em coluna única, desconsiderando a formatação em duas ou mais colunas, foi adequada e necessária. Apenas uma biblioteca não apresentou esse resultado. De fato, houve uma diferença significativa na

fidelidade terminológica, fato norteador na escolha da biblioteca de conversão. Ainda assim, duas bibliotecas ofereceram resultados aceitáveis para o processamento das consultas.

Com relação ao que foi denominado aqui de “resultados estatísticos”, percebeu-se, a partir dos dados obtidos, que a terminologia de maior revocação foi a MeSH. Isso foi verificado ainda que a MeSH possua princípios tradicionais em sua organização que a tornam menos flexível, por exemplo, nas relações entre os termos, além do fato de que contém menor número de termos em relação a SNOMED CT. Nesse sentido, o experimento usou 56.130 termos da SNOMED e 54.582 termos da MeSH.

Os resultados foram, de certa forma, surpreendentes. A constatação que uma terminologia com estruturação hierárquica mais rígida e com menor número de termos na mesma área de conhecimento obteve maior revocação no experimento prático, deixa em aberto a reflexão sobre os princípios que podem nortear os sistemas de organização do conhecimento e sua aplicação em mecanismos de recuperação da informação. Com relação às primeiras considerações a respeito desses resultados pode-se elencar:

i. O número de termos no recorte (*diseases*) na terminologia SNOMED CT foi relativamente equilibrado em relação à terminologia MeSH. Mesmo considerando de forma geral, a quantidade muito maior de termos para a totalidade da SNOMED CT.
ii. Pelo número de termos na expansão de consultas, o tempo de processamento para a SNOMED CT é, em média, um segundo maior. Esse custo computacional deve ser levado em conta em aplicações de tempo real, visto que pode ser um fator limitador na implantação de sistemas *online*.
iii. O grande número de termos associados ao termo principal em consultas na SNOMED CT não representou o mesmo impacto na revocação.
iv. Observa-se uma tendência de que o número de palavras nos termos de cada terminologia: enquanto a MeSH trabalha com termos mais curtos, a SNOMED CT faz uso de descrições maiores.

Pode-se inferir que não apenas um, mas sim um conjunto de fatores foi responsável pelos resultados obtidos. A notar pelos primeiros indicadores estatísticos, a terminologia que a princípio possuía mais recursos – maior número de termos e mapeamento dinâmico pelas conexões axiomáticas – obteria maior revocação dos documentos científicos, mas os resultados não demonstraram isso.

Apesar da SNOMED CT, a partir dos termos comuns entre as terminologias, acumular 1.548 termos a mais que a MeSH, a revocação correspondente não foi influenciada por sua superioridade numérica. Os dados revelaram que para cada consulta submetida, em

média, a SNOMED CT tinha um termo a mais que a MeSH, e mesmo assim não obteve melhores dados quantitativos.

Interessante notar também que, mesmo procurando alinhar o tema dos artigos e o grupo de termos comuns selecionados nas terminologias, não houve nenhuma revocação para um número expressivo de termos. Isso se verificou considerando-se como critério a ausência de revocação ao mesmo tempo em ambas terminologias para a mesma pesquisa comum. Do total de 2.225 termos comuns, 1.387 termos (e seus conjuntos) não retornaram nenhum artigo, levando em consideração a ausência de revocação em ambas terminologias no mesmo processo. Em análise oposta, 698 termos (e seus conjuntos) retornaram revocação no mesmo processo, ou seja, no momento que o termo comum e seus termos associados são pesquisados em cada terminologia.

Essa reflexão revela que, mesmo procurando alinhar a temática com o acervo de artigos, permanece um distanciamento entre os termos dos artefatos e a linguagem autoral. Esse distanciamento pode se originar nas diferentes concepções dos especialistas dos sistemas de representação e dos autores de artigos científicos, bem como na falta de consulta às terminologias ao escrever artigos. Tal falta de alinhamento cria dificuldades nos processos de RI por correspondência sintática, ainda que o mesmo conceito possa ser representado em sintaxes diferentes.

Uma outra observação, que concorda com a percepção de opinião comum, é que o número de palavras na representação do termo impacta na RI. Esse aspecto foi percebido com clareza no experimento realizado: a SNOMED CT, quando não houve revocação de artigos, apresentou em média, termos com 4,38 palavras; na mesma situação, a MeSH apresentou 1,93 palavras. Houve equilíbrio ao identificar a média de palavras nos termos quando havia revocação: 1,93 palavras por termo para SNOMED CT e 1,94 palavras por termo na MeSH. A inclusão de termos compostos e complexos dificulta o processo de RI. Isso indica a maior probabilidade de revocação quando os termos possuem menos palavras, ou seja, o potencial de correspondência textual é afetado diretamente por esta métrica. Ainda em relação à métrica do quantitativo de palavras, houve medição para os termos comuns entre as terminologias, com uma média ainda menor, de 1,73 palavras por termo.

A respeito dos termos comuns que nortearam a expansão de consultas, foi realizado um experimento onde um algoritmo semelhante ao usado no pré-processamento de artigos foi aplicado no grupo *diseases* a partir da MeSH. Esse conjunto de termos tratado com a exclusão de caracteres especiais, números, adequação de espaços, entre outras técnicas, foi submetido a correspondência com a SNOMED CT. O resultado foi um acréscimo de aproximadamente 500

termos em relação à correspondência anterior dos termos originais, sem tratamento. Isso serviu de parâmetro para sinalizar que as terminologias, mesmo apontando para os mesmos conceitos, possuem particularidades de sintaxe que podem dificultar iniciativas de interoperabilidade e também de representação comum. Em geral os tesauros são criados *ad-hoc*, com recorte e abordagem específicos.

Como os termos comuns foram baseados nos descritores da terminologia MeSH, foram feitas análises específicas. Uma delas destaca que, quando não há revocação pelos termos associados, em 602 consultas, a MeSH obtém maior revocação apenas pelo termo comum. Por outra análise, considerando apenas a revocação de termos associados, em 149 consultas, a MeSH conseguiu recuperar artigos, ao passo que em 93 consultas a SNOMED CT conseguiu revocação. Desse modo, mesmo considerando o termo comum como o representante conceitual centralizador (na MeSH), houve momentos onde a revocação se deu exclusivamente pelos termos associados. Mais uma vez, a MeSH destacou-se em comparação com a SNOMED CT.

Quando se trata do somatório de artigos em revocação, de forma ampla, ou seja, incluindo artigos repetidos na mesma consulta, os números são ainda mais expressivos e demonstram a maior capacidade de correspondência da MeSH do que da SNOMED CT. Nesta análise foram 21.557 artigos recuperados no total pela MeSH, em comparação com 7.068 artigos recuperados pela SNOMED CT. Uma diferença de 14.489 artigos, ou seja, a SNOMED CT recuperou aproximadamente um terço em comparação com a MeSH, partindo dos mesmos termos comuns. Essa análise geral ajuda a demonstrar que os termos da MeSH possuem maior possibilidade de correspondência em relação aos artigos científicos.

Mesmo quando se elimina a revocação de artigos repetidos, artigos que podem ser recuperados mais de uma vez por termos diferentes, ainda assim a terminologia MeSH alcança quase o dobro em revocação, com 9.693 artigos versus 5.835 artigos da SNOMED CT. Ainda, na hipótese de uma terminologia não recuperar nenhum artigo e a outra terminologia conseguir revocação, a análise também exibe vantagem para a MeSH: são 98 consultas onde a SNOMED CT não retorna nenhuma revocação e a MeSH consegue recuperar artigos, contra 42 situações inversas.

Uma análise importante foi identificar na média de consultas onde houve revocação o número de termos usados em cada terminologia. Foram 42.852 termos usados na MeSH e 41.485 termos da SNOMED CT. Com esse conjunto, cada terminologia recuperou 21.557 e 7.068 artigos únicos, respectivamente, conforme já citado. Esses números permitem inferir que mesmo com a proximidade do uso de termos em ambas terminologias, o número de artigos recuperado é díspar, mantendo-se a vantagem para a MeSH.

Uma comparação final se refere ao tempo gasto em segundos para processar cada conjunto de consulta. Como foi observado anteriormente, o fato da SNOMED CT possuir mais termos repercutiu diretamente nessa análise: o tempo no processamento da consulta foi em média um segundo maior que na MeSH. É válido avaliar que, independentemente da terminologia, a demora no processo de RI pode impactar na viabilidade de implementação de soluções em tempo real, onde o usuário anseia por respostas rápidas. Isso ocorre principalmente em ambientes web, multiusuário, gerando desafios para computação concorrente no envio e recepção de dados para a interface do usuário.

Diante da observação e análise dos dados obtidos pelo processo de submissão de consultas expandidas, acredita-se que o destaque da MeSH está relacionado à sua origem e objetivos iniciais. Conforme o processo histórico de cada terminologia, a MeSH tem como bases listas de autoridade indicadas para a representação documental, enquanto a SNOMED CT objetiva a representação clínica de registros médicos. Essa diferença na abordagem conceitual foi traduzida nos resultados encontrados no experimento.

7 CONSIDERAÇÕES FINAIS

Este trabalho foi planejado para comparar dois artefatos de representação do conhecimento, em suas aplicações práticas em recuperação da informação pelo método de expansão de consultas. Para realizar uma pesquisa de caráter aplicado, foi desenvolvido um *software* capaz de manipular as terminologias e os artigos, além de criar um ambiente passível de submissão de consultas expandidas por estruturas hierárquicas, sinônimos e termos conceitualmente próximos. A própria criação desse ambiente demandou um conjunto de processos e técnicas, antes de propriamente chegar às consultas. Um modelo sequencial de etapas foi apresentado (seção 5.2.2.2, figura 4) que organizou em grupos os principais processos do experimento.

No capítulo 1, o tema foi a própria Recuperação da Informação, histórico e principais características. Pesquisou-se as motivações para RI, sua evolução como área científica e os autores que fizeram grande contribuição. Identificar a evolução pelas décadas passadas foi uma pesquisa prazerosa, em que foi possível perceber a proximidade da evolução das terminologias médicas frente aos desafios da RI em ambiente digital. Ao adentrar na RI direcionada às terminologias, encontram-se os Sistemas de Organização do Conhecimento, estruturas que suportam a representação em diversas formas e que permitem interações para diversos propósitos, inclusive RI.

No capítulo 2, a atenção voltou-se para dois tipos de SOCs: os tesauros e as ontologias. Esses modelos tipificam as terminologias médicas usadas no trabalho: o MeSH tem sua estrutura baseada nos princípios dos tesauros, e a SNOMED CT utiliza constructos formais, axiomas em lógica descritiva na definição de termos e relações. Cada terminologia recebe atenção em seções distintas, explorando o processo histórico, principais aspectos estruturais e aplicações em domínios específicos. Observou-se na MeSH uma intrínseca conexão com a representação documental e suas características conceituais. Na SNOMED CT, observaram-se suas características e princípios que nortearam perfil do artefato na direção do registro eletrônico médico. Os pilares a) conceitos, b) descrições e c) relacionamentos formam a base estrutural da terminologia que possui mecanismos diferenciados de organização, com o uso de constructos formais, a saber, axiomas para definir os termos e suas relações. Essa estrutura possibilita ainda o uso de raciocinadores para inferência automática.

O capítulo 3 abordou os paradigmas de banco de dados, a expansão de consultas e seus aspectos de aprimoramento da consulta original, adicionando termos com o objetivo de aumentar a revocação.

O capítulo 4 apresenta a metodologia que norteou a pesquisa. Os passos metodológicos incluem: a) a aquisição das terminologias e seus aspectos administrativos de licenciamento e *download*; b) a estruturação terminológica para um modelo de entidade-relacionamento, apto a responder consultas aos termos, bem como recuperar termos sinônimos e associados; c) a aquisição de artigos, em formato completo e de publicação livre, a fim de constituir um acervo na mesma área de conhecimento das terminologias; d) pré-processamento de artigos que envolve a extração de texto do formato proprietário originado do site publicador, além da adequação textual e eliminação de caracteres especiais, dentre outros processos de preparação; e) a estruturação dos artigos em um banco de dados capaz de responder por pesquisas texto completo*;* f) o desenvolvimento do experimento fazendo uso dos recursos já elencados, submissão de *consultas* respeitando características das terminologias para expansão e armazenamento de resultados em um banco de dados para análise.

O capítulo 5 trouxe resultados e as discussões. Para cada conjunto de passos da metodologia, apresentou-se reflexão correspondente sobre os resultados. Abordou-se assim, em mais detalhes, cada etapa da metodologia para em seguida tecer considerações técnicas e práticas, buscando compreender os detalhes de cada processo. Nas discussões, há um direcionamento dos resultados do trabalho que reflete sobre as limitações e os problemas encontrados. Foi possível ainda discutir o resultado do questionário sobre terminologias clínicas divulgado na comunidade acadêmica e profissional e observar a percepção desse público sobre esses sistemas de organização de conhecimento e suas aplicações. Em outra seção, a discussão sobre os resultados da pesquisa discorreu sobre a compreensão dos dados de resultado, confirmando a maior revocação da terminologia MeSH e observando suas características em comparação com a SNOMED CT, a fim de elencar princípios para trabalhos de outros pesquisadores e para trabalhos futuros.

Enfim, os resultados obtidos produziram direcionamentos que permitem reafirmar a importância de sistemas de organização do conhecimento, especificamente, as terminologias biomédicas na recuperação da informação. Na seção seguinte, a Seção 7.1, evidenciam-se características terminológicas e seus impactos; a Seção 7.2 se atém ao experimento com o *software*; e finalmente, a seção 7.3 conclui sobre terminologia e RI, ressalta as contribuições e apresenta alternativas de trabalho futuro.

7.1 Características terminológicas

As características terminológicas discutidas aqui envolvem o histórico, a quantidade de termos por terminologia, as diferentes relações terminológicas e a quantidade de palavras por termo.

Com relação ao histórico, observou-se que ambos artefatos terminológicos tiveram origem em outras iniciativas, que foram se mesclando e se combinando não se sabe bem de que forma (vide Seção 3.2). Isso ajuda a entender características que afetaram o experimento, como por exemplo, a falta de padrão na sintaxe dos termos e a própria formatação de caracteres. A adição de diversas terminologias no passado para compor a estrutura terminológica da MeSH e SNOMED CT contribuíram para a caracterização atual, onde observa-se a necessidade de normalização de termos e de sua forma sintática.

A perspectiva de Schütze e Pedersen (1997), sugere que tesauros são úteis para RI quando oferecem sinônimos de termos usados na definição do corpus. Essa abordagem é válida, mas não é um princípio singular.

O presente trabalho concorda com o entendimento de Dextre (2016), em que instrumentos terminológicos não teriam mais espaço entre os sistemas de recuperação modernos, amplos e abrangentes, a exemplo de buscadores web. Ainda assim, são importantes ferramentas em domínios específicos, onde podem exercer importante papel na expansão de consultas e no controle de vocabulário. Corrobora também com a visão de Svenonius (1986), que esses artefatos atuam, portanto, para mitigar problemas linguísticos de ambiguidade e mediar a representação entre a linguagem do usuário e autor.

A relevância das terminologias no presente trabalho não esteve atrelada à processos de indexação ou categorização, mas na expansão para a recuperação da informação. As relações semânticas foram importantes para o aumento de consultas, mas não foi fator decisivo para os resultados. Acrescenta-se o fato que não se avaliou aqui a precisão dos documentos alinhados ao interesse da consulta, mas sim sua revocação.

Com relação à quantidade de termos por terminologia, destaca-se que tal característica influenciou a visão pré-concebida no início da pesquisa. Conforme Seções 5.1.6.4 e 5.2.3.4, a quantidade de termos identificada foi:

Tabela 2 - Quantidade de termos por terminologia

Termos	**MeSH**	**SNOMED CT**
Total geral	179.989	1.216.589
Comuns + associados usados no algoritmo	56.807	58.355

Fonte: Dados da pesquisa

Houve expectativa inicial de que a vantagem quantitativa da SNOMED CT resultasse em superioridade sobre a MeSH, fato que não ocorreu conforme a análise de dados demonstrou. Outra questão, levantada por alguns autores (vide Seção 1), seria uma suposta vantagem das ontologias na RI, ou seja, das terminologias que possuem constructos formais sobre o uso de sistemas de organização do conhecimento que não possuem tais características, como os tesauros. Essa suposição não pôde ser evidenciada pelo experimento prático desse trabalho, ainda que por alguns autores, a SNOMED CT seja considerada uma ontologia (não há consenso sobre essa classificação), mesmo que suas relações sejam baseadas em axiomas. O que pôde-se notar é que mesmo a SNOMED CT possua um maior número de termos, sua revocação foi inferior à MeSH.

Com respeito às diferentes formas de relações terminológicas, notou-se no experimento que a capacidade de estabelecer vínculos entre os termos é importante, mas também não parece ser um fator determinante, tomado de forma isolada. Ambas terminologias partem de conceitos principais para: a) identificar descrições diferenciadas sobre o mesmo conceito, b) identificar termos associados a um conceito principal, c) identificar estruturas hierárquicas relacionadas ao conceito. Nestas relações ressaltam-se características da terminologia SNOMED CT:

- A partir de uma *string* (termo), identifica-se o *concept id*;
- A partir desse *id*, procura-se os termos associados e os termos hierárquicos;
- Os termos hierárquicos são definidos por axiomas em OWL nas seguintes expressões: *ObjectIntersectionOf*; *EquivalentClasses*; *SubClassOf*; de forma recursiva, até que toda estrutura seja avaliada.

Ainda a respeito das relações terminológicas, no caso da MeSH, destaca-se:

- A partir de uma *string* (termo), identifica-se o *descriptor ui*;
- A partir desse *ui*, procura-se os termos de entrada e os termos hierárquicos;
- Os termos hierárquicos são definidos por codificação rígida, em numeração organizada por blocos, que após a estruturação são recuperados por SQL no banco de dados.

Essas características diferenciadas nas relações terminológicas foram percebidas em algumas consultas com uma disparidade que merece atenção. Adiciona-se a esse contexto, o fato de que nem todas as expressões disponíveis nos axiomas da terminologia SNOMED CT foram usadas, como por exemplo: *ObjectSomeValuesFrom*. De forma similar, as possibilidades de expansão da terminologia de forma automática por raciocinadores (*reasoners*) também não foi aproveitada. A seção 5.1.2.2 exibe um conjunto de análises de resultados que apoia a compreensão e o impacto das diferentes formas de relações terminológicas. Acrescenta-se aqui, como consideração adicional, os 20 primeiros quantitativos de termos usados para consultas em ordem decrescente.

Com relação à SNOMED CT, observa-se que os 5 primeiros termos comuns deram origem a uma grande quantidade de consultas. Porém, desses termos, propriamente, a quantidade de termos com revocação é baixa (tabela 3). Embora a terminologia use axiomas para as conexões e que tal característica permita boas possibilidades de conexões adicionais, a revocação é baixa para um alto custo computacional, como demonstrou o experimento.

Tabela 3 - Os 20 primeiros termos da SNOMED CT em quantidade decrescente

Termo comum	Quant Termos	Termos com Revocação	Artigos Recuperados	Revocação
poisoning	2656	9	49	1,98%
pain	2633	119	485	19,55%
drug overdose	2129	1	7	0,28%
intellectual disability	1283	7	26	1,05%
anemia	1058	17	44	1,77%
malnutrition	776	38	143	5,76%
epilepsy	748	12	31	1,25%
hearing loss	733	22	85	3,43%
eczema	580	5	11	0,44%
lymphadenopathy	530	2	2	0,08%
strabismus	495	2	4	0,16%
glaucoma	416	1	10	0,40%
ataxia	412	3	12	0,48%

spasm	367	5	10	0,40%
dementia	362	27	135	5,44%
nephrosis	356	5	33	1,33%
diabetes mellitus	333	12	55	2,22%
acute disease	317	16	82	3,31%
diarrhea	308	7	26	1,05%
sepsis	286	4	17	0,69%

Fonte: Dados da pesquisa

Em comparação, embora a MeSH (tabela 4) tenha em seus dois primeiros termos comuns uma quantidade de termos díspar do restante da coleção, nota-se nesse conjunto das 20 maiores quantidades de termos usados para consultas, que a quantidade de termos com revocação é mais alta que na SNOMED CT. Pode-se inferir que a MeSH foi mais eficiente para representar a linguagem dos autores e o número de artigos recuperados é outro indicativo desse resultado.

Tabela 4 - Os 20 primeiros termos da MeSH em quantidade decrescente

Termo comum	**Quant. Termos**	**Termos com Revocação**	**Artigos Recuperados**	**Revocação**
nervous system diseases	10060	576	2930	118,10%
neoplasms	4606	212	765	30,83%
digestive system diseases	1626	217	1046	42,16%
heart diseases	1029	118	693	27,93%
parasitic diseases	703	14	50	2,02%
lung diseases	580	70	369	14,87%
connective tissue diseases	559	17	108	4,35%
kidney diseases	545	50	266	10,72%
ear diseases	461	36	188	7,58%
carcinoma	445	29	110	4,43%

epilepsy	436	13	46	1,85%
hypersensitivity	418	14	79	3,18%
pregnancy complications	402	11	16	0,64%
lipidoses	356	1	1	0,04%
helminthiasis	347	1	1	0,04%
poisoning	329	12	50	2,02%
pain	327	68	486	19,59%
adenocarcinoma	313	13	51	2,06%
leukemia	308	9	30	1,21%
mycoses	287	15	28	1,13%

Fonte: Dados da pesquisa

As relações terminológicas são, sem dúvida, um importante fator na caracterização dos sistemas de organização do conhecimento. Nesse aspecto, apenas os 20 primeiros termos comuns que geraram maior associação terminológica foram considerados.

Com relação a quantidade de palavras por termo, trata-se de uma característica marcante que impactou diretamente na revocação. Expressar o mesmo conceito com poucas palavras, foi um fator diferencial. A seção 6.1.2.2 exibe a média de quantidade de palavras por termos, destacando a terminologia SNOMED CT com uma média de quase o dobro de palavras por termo, em comparação com a MeSH.

Acredita-se que o processo de criação de cada terminologia exerceu grande influência em seu perfil descritivo. Enquanto a terminologia MeSH nasce em um contexto de representação documental, a terminologia SNOMED CT é direcionada para outro foco, na descrição de registros eletrônicos de saúde em diagnósticos clínicos. Essa diferença foi notada nesse experimento em RI frente aos artigos científicos, e permite inferir que terminologias, ainda que adotem aspectos estruturais mais simples, com menor número de termos, mas dotada de sintaxe com menor número de palavras na capacidade de expressão conceitual, pode ser mais adequada a processos de RI. O aumento da formalização nem sempre é benéfico nesse cenário, especificamente.

7.2 Experimento em software

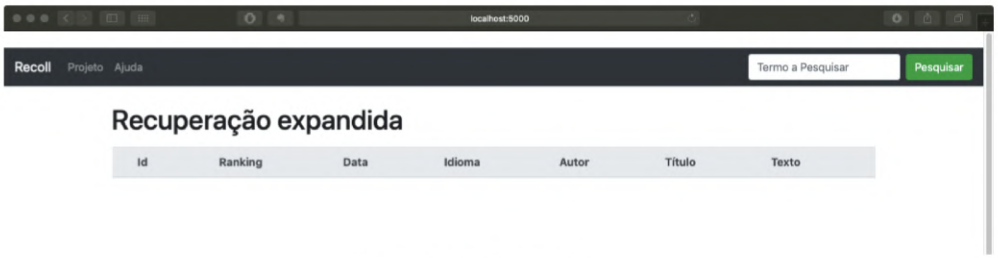
Houve um alto dispêndio de tempo na construção do *software*, o que impactou diretamente no desenvolvimento do trabalho, mesmo com a opção de utilizar uma linguagem de programação produtiva e de alto nível. Ainda, as alterações no projeto de pesquisa, desafio inerente a qualquer pesquisador, e as novas descobertas durante o desenvolvimento do sistema, trouxe um esforço adicional para concluir o trabalho. Da forma como foi estruturado, o trabalho escrito dependeu dos resultados do experimento para ser desenvolvido.

Registra-se ainda que duas grandes funcionalidades foram desenvolvidas, mas não implementadas nesse momento para cumprir os prazos do programa da pós-graduação:

a) *Resultados com termos normalizados*: da mesma forma que o texto dos artigos, os termos foram normalizados no pré-processamento. Em testes locais, após o tratamento, aproximadamente 500 outros termos obtiveram correspondência entre as duas terminologias, formando um novo grupo comum. Esse acréscimo certamente deve impactar nos resultados de revocação, da mesma forma que impactaram os termos associados. A comparação desses dois processos (com termos originais e termos normalizados) era algo planejado, mas que será usado na continuidade da pesquisa (ver Seção 7.3 adiante). O algoritmo já se encontra parametrizado para realizar as consultas pelo termo original ou pelo termo normalizado. O banco de dados, também está preparado para tal modelo, com colunas separadas para as descrições originais e tratadas (normalizadas).

b) *Interface de alto nível para consulta em tempo real*: a fim de criar um mecanismo para demonstrar ao usuário final o funcionamento do processo de expansão de consulta foi criada uma interface via browser. A ideia era permitir a interação com o algoritmo e um melhor entendimento por alunos, pesquisadores e profissionais interessados nesta temática. Os algoritmos foram criados prevendo este tipo de uso futuro, conforme a figura 4 e também a figura 42. A interface web funciona baseada no micro servidor *Flask* e provê acesso simples e funcional para a submissão de consultas. A figura 32 abaixo ilustrar a interface:

Figura 30 – Interface software SRAM para usuário final



Fonte: Dados da pesquisa

Embora em sua maior parte descritiva, o presente trabalho tenha destacado o *software* que processou as *consultas* e gerou resultados para análise (Seção 4.2.2.2), o desenvolvimento do experimento como um todo, exigiu a criação de outros *softwares* independentes com as seguintes funções:

a) Download automatizado (*scraping*) dos 2.481 artigos da *BMC Geriatrics*;
b) Extração de texto e metadados, e envio de artigos para o banco de dados (*elasticsearch*);
c) Criação de bancos de dados (*script*), tratamento e importação das terminologias;
d) Identificação de termos comuns entre as terminologias, a partir do grupo *diseases*.

Como opinião pessoal, destaca-se que desenvolvimento do experimento em *software* se mostrou um processo desafiador, mas gratificante. A forma de trabalho agregou conhecimento, além de permitir ao pesquisador conhecer tecnologias e procurar soluções para problemas comuns da RI, reconhecidos por diversos autores de forma teórica. A prática, portanto, permitiu endossar certas teorias, bem como ampliar a discussão sobre outras.

Acredita-se que a forma de pesquisa aplicada proporcionou melhor entendimento sobre o uso de terminologias na expansão de consultas, bem como seus efeitos práticos na recuperação de documentos científicos. O aumento de revocação foi evidente e aponta a viabilidade da técnica em ambientes específicos, mas não de modo não generalizado, o que corrobora o caráter qualitativo da pesquisa. Há também a percepção que o aumento da revocação não garante a qualidade da recuperação, para suportar tal afirmação seria necessário um experimento com validação de especialistas da área. Mas o aumento de revocação permite maiores possibilidades de identificação de respostas ao usuário. Portanto, acredita-se na utilidade do experimento na contribuição para aplicação prática, demonstrando etapas e processos importantes nesse modelo de RI. Ao *software* foi dado o nome de Sistema de recuperação de artigos médicos por expansão de *queries* em artefatos terminológicos (SRAM).

7.3 Contribuições e trabalhos futuros

Nessa seção, além de contribuições e trabalhos futuros, destacam-se considerações finais sobre terminologias e RI.

Como principais contribuições desta pesquisa, pode-se citar:

a) A possibilidade de utilização do *software* para demonstrações prático-pedagógicas, em ambiente educacional, de forma a demonstrar aos discentes as características e etapas do processo de RI baseado em expansão de consultas por terminologias;
b) A utilização do *software* para comparação de terminologias, no sentido de identificar princípios que podem nortear o trabalho de construção e manutenção de terminologias;
c) A percepção de todo um conjunto de etapas necessárias que antecederam a submissão de consultas, à saber, a preparação do ambiente, a aquisição de terminologias, a estruturação de dados, o pré-processamento de artigos, bem como o uso de diversas bibliotecas e estruturas de *software* para implementação;
d) A publicação do *software* com registro no INPI, em repositório digital livre, para promover o acesso a discentes e pesquisadores, ampliando as discussões sobre a aplicação de terminologias em RI;
e) A demonstração que para RI, a construção da terminologia em sua representação terminológica possui maior impacto em comparação com as possibilidades de conexões e quantidade de termos.

Como possibilidades **futuras**, considera-se:

a) A reformulação do *software* para aceitar outros formatos de representação terminológicas, como RDF e OWL, além de permitir terminologias de outras áreas de pesquisa;
b) A integração do *software* com outras soluções tecnológicas de construção de sistemas de organização do conhecimento, a exemplo da ferramenta Onto4ALL[49] Editor, onde este modelo de RI por consulta expandida pode fazer uso de artefatos ali criados;
c) A construção de banco de dados de documentos científicos de áreas diversas para permitir o uso de pesquisas empíricas, por outras áreas de conhecimento;
d) A conclusão da interface de alto nível (Figura 43) para uso por não especialistas da computação e capaz de:

[49] https://onto4alleditor.com/pt

- Permitir a submissão de consultas em linguagem natural, identificando termos passíveis de expansão pela(s) terminologia(s) adicionada(s) (algoritmo pronto).
- Permitir a adição de novas terminologias;
- Possibilitar um *log* de etapas, para fins didáticos, ilustrando para o usuário os processos envolvidos em SRIs.
- Gerar gráficos estatísticos semelhantes aos aqui apresentados, para ajudar na compreensão das possibilidades de recuperação.

Finalmente, cabem ainda algumas considerações finais sobre terminologia e RI. Foi possível perceber no presente trabalho, usando o experimento prático como pesquisa aplicada, que não há uma decisão simples, única, a respeito de qual tipo de SOC é o ideal para RI considerando os critérios adotados: um ambiente específico, controlado, em determinada área do conhecimento. A comparação dos artefatos possibilitou deduzir, de forma qualitativa, que:

- Em termos de estrutura terminológica: SOCs que utilizam constructos formais, como as ontologias, trazem alternativa de representação superiores às estruturas tradicionais como os tesauros.
- Em termos de sintaxe terminológica: SOCs que utilizam conceitos representados de forma abrangente, criados para fins de indexação e categorização, são mais adequados à RI pois a estratégia indica maior proximidade da linguagem autoral.

Portanto, pôde-se perceber que a escolha do tipo de terminologia para RI não é simples, ou baseada apenas no tipo de SOC de forma isolada. Reconhece-se que a representação sintática precede a estruturação terminológica em termos de diferencial para RI, e a combinação desses dois fatores é o que pode potencializar a expansão de consultas em sistemas de RI, visando resultados satisfatórios ao usuário final.

REFERÊNCIAS

ABDOU, Samir; SAVOY, Jacques. Searching in Medline: Query expansion and manual indexing evaluation. **Information Processing & Management**, v. 44, n. 2, p. 781–789, mar. 2008.

ALLEN, Bryce. Topic Knowledge and Online Catalog Search Formulation. **The Library Quarterly**, v. 61, n. 2, p. 188–213, abr. 1991.

ALMEIDA, Mauricio Barcellos. 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, Maurício Barcellos. Uma abordagem integrada sobre ontologias: Ciência da Informação, Ciência da Computação e Filosofia. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 19, n. 3, p. 242–258, set. 2014. https://doi.org/10.1590/1981-5344/1736.

ALMEIDA, Maurício Barcellos; AGANETTE, Elisângela Cristina. Terminologia e ontologia: discussões sobre a criação de definições em vocabulários biomédicos. **Ciência da Informação**, v.45, n.1, p.11-24, 2017.

ALMEIDA, Maurício Barcellos; MENDONÇA, Fabrício Martins; AGANETTE, Elisângela Cristina. Interfaces entre ontologias e conceitos seminais da Ciência da Informação: Em busca de avanços na organização do conhecimento. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO (ENANCIB), XIV, 2013, **Anais...** Marília: UNESP, 2013 p.22. (GT 2 - Organização e Representação do Conhecimento Comunicação Oral). Disponível em http://mba.eci.ufmg.br/downloads/OntologiasAltoMedioNivel_Enacib2013_camera_ready.pdf. Acesso em 09 out 2020.

ARONSON, Alan R; RINDFLESCH, Thomas C. Query Expansion Using the UMLS® Metathesaurus®.In: **Proceedings of the AMIA Annual Fall Symposium**, 1997, p. 485-9.

AZAD, Hiteshwar Kumar; DEEPAK, Akshay. Query expansion techniques for information retrieval: A survey. **Information Processing & Management**, v. 56, n. 5, p. 1698–1735, set. 2019.

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

BERNERS-LEE, Tim.; HENDLER, James; LASSILA, Ora. The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. **Scientific American**, p. 29-37, Jun 2006. Disponível em: http://link-galegroup.ez27.periodicos.capes.gov.br/apps/doc/A147823458/AONE?sid=googlescholar. Acesso em: 14 abr. 2018.

BERNIER, C. L.; CRANE, E. J. Indexing abstracts. **Industrial and Engineering Chemistry**. v.40, n.4, p. 725–30, 1948.

BHATTACHARYYA, S. B. SNOMED CT History and IHTSDO. In: BHATTACHARYYA, S. B. **Introduction to SNOMED CT**. Singapore: Springer, 2016. p. 19–23. Disponível em: http://link.springer.com/10.1007/978-981-287-895-3_3. Acesso em: 17 jun. 2020.

BJÖRK, Bo-Christer; ROOS, Annikki; LAURI, Mari. Global annual volume of peer reviewed scholarly articles and the share available via different Open Access options.In: **Proceedings ELPUB2008 Conference on Electronic Publishing,** Toronto, Canada, June 2008, p. 11.Disponível em: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.162.991&rep=rep1&type=pdf. Acesso em 09 out 2020.

BORGMAN, Christine L. All users of information retrieval systems are not created equal: An exploration into individual differences. **Information Processing & Management**, v. 25, n. 3, p. 237–251, jan. 1989.

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

BRITO, Ricardo Wagner. **Bancos de Dados NoSQL x SGBDs Relacionais**: Análise Comparativa. p. 1-6, 2010. Disponivel em http://docplayer.com.br/433629-Bancos-de-dados-nosql-x-sgbds-relacionais-analise-comparativa.html. Acesso em 08 out 2020.

BUSH, Vannevar. As We May Think. **The Atlantic**, jul. 1945. Disponível em: https://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/. Acesso em: 14 abr. 2018.

CARPINETO, Claudio; ROMANO, Giovanni. A Survey of Automatic Query Expansion in Information Retrieval. **ACM Computing Surveys**, v. 44, n. 1, p. 1–50, 1 jan. 2012.

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

CLARKE, Stella G. Dextre. **Knowledge Organization,** v. 43, n. 3, p. 138–144, 2016.

CLEVERDON, Cyril. Evaluation Tests of Information Retrieval Systems. **Journal of Documentation**, v. 26, n. 1, p. 55–67, jan. 1970.

CLEVERDON, Cyril; MILLS, Jack; KEEN, Michael. **Factors determining the performance of indexing systems**. Indiana:College of Aeronautics, 1966.

CIMINO, James J. Vocabulary and health care information technology: State of the art. **Journal of the American Society for Information Science**, v.46, n.10, p.777-782, dec 1995.

CODD, E F. A Relational Model of Data for Large Shared Data Banks. **Communications of the ACM**, v. 13, n. 6, p. 11, 1970.

COOL, Colleen; BELKIN, Nicholas J. Interactive information retrieval: history and background. In: RUTHVEN, Ian; KELLY, Diane. (Eds.). **Interactive Information Seeking, Behaviour and Retrieval**. [*S. l.*]: Facet, 2013. p. 1–14. Disponível em: https://www.cambridge.org/core/product/identifier/CBO9781856049740A011/type/book_part. Acesso em: 30 mar. 2020.

CURRÁS, Emilia. **Tesauros, linguagens terminológicas**. Brasília: IBICT, 1995. Disponível em: http://livroaberto.ibict.br/handle/1/454. Acesso em: 6 jun. 2018.

EFTHIMIADIS, Efthimis N. Interactive query expansion: A user-based evaluation in a relevance feedback environment. **Journal of the American Society for Information Science**, v.51, n.11, p. 989–1003, 2000.

EL-SAPPAGH, Shaker; FRANDA, Francesco; ALI, Farman; KWAK, Kyung-Sup. SNOMED CT standard ontology based on the ontology for general medical science. **BMC Medical Informatics and Decision Making**, v. 18, n. 1, p. 76, dez. 2018.

ELMASRI, Ramez; NAVATHE, Sham. **Fundamentals of database systems**. 6th ed. Boston: Addison-Wesley, 2011.

FARINELLI, Fernanda; ALMEIDA, Mauricio Barcellos; ELKIN, Peter; SMITH, Barry. OntONeo: The Obstetric and Neonatal Ontology. p. 7, 2016.

FELBER, Helmut. **Terminology Manual**. Vienna (Austria): International Information Centre for Terminology, 1984. 483p. Disponível em: https://eric.ed.gov/?id=ED254245. Acesso em: 7 jul. 2020.

FRICKÉ, Martin. **Logic and the Organization of Information**. New York, NY: Springer New York, 2012. Disponível em: http://dx.doi.org/10.1007/978-1-4614-3088-9. Acesso em: 18 jan. 2019.

GILCHRIST, Alan. Thesauri, taxonomies and ontologies – an etymological note. **Journal of Documentation**, v. 59, n. 1, p. 7–18, fev. 2003.

GIUNCHIGLIA, Fausto; DUTTA, Biswanath; MALTESE, and Vincenzo. From Knowledge Organization to Knowledge Representation. **Knowledge Organization**, v. 41, n. 1, p. 44–56, 2014.

GUELPELI, Marcus Vinicius Carvalho. **Cassiopeia**: Um modelo de agrupamento de textos baseado em sumarização. 2012.221f. Tese (Doutorado em Computação- Inteligência Artificial) - Universidade Federal Fluminense, Niterói, 2012.

HAJBA, Gábor László. **Website Scraping with Python**: Using BeautifulSoup and Scrapy. Berkeley, CA: Apress, 2018. Disponível em: http://link.springer.com/10.1007/978-1-4842-3925-4. Acesso em: 3 maio 2020.

HERSH, William R. **Information retrieval: a health and biomedical perspective**. 3rd ed. New York, NY: Springer, 2009. (Health informatics).

HJØRLAND, Birger. Does the Traditional Thesaurus Have a Place in Modern Information Retrieval? **Knowledge Organization**, v. 43, n. 3, p. 145–159, 2016.

HOLLINK, Laura; MALAISÉ, Véronique; SCHREIBER, Guus. Thesaurus enrichment for query expansion in audiovisual archives. **Multimedia Tools and Applications**, v. 49, n. 1, p. 235–257, ago. 2010.

HOPPE, Thomas; HUMM, Bernhard; REIBOLD, Anatol. **Semantic applications**. New York, NY: Springer, 2018.

JINHA, Arif E. Article 50 million: an estimate of the number of scholarly articles in existence. **Learned Publishing**, v. 23, n. 3, p. 258–263, jul. 2010.

KELECHAVA, Brad. The SQL Standard - ISO/IEC 9075:2016 (ANSI X3.135). **American National Standards Institute - ANSI Blog**. October 5, 2018. Disponível em: https://blog.ansi.org/2018/10/sql-standard-iso-iec-9075-2016-ansi-x3-135/. Acesso em: 4 fev. 2020.

LACY, Lee W. **OWL:** representing information using the web ontology language. Victoria, B.C: Trafford Publishing, 2005.

LARA, M. L. G. de. O unicórnio (o rinoceronte, o ornitorrinco...), a análise documentária e a linguagem documentária. **DataGramaZero**, v.2, n.6,2001.

LEMOS, Daniela Lucas da Silva; SOUZA, Renato Rocha. Knowledge Organization Systems for the Representation of Multimedia Resources on the Web: A Comparative Analysis. **Knowledge Organization**, v.47, n.4, p. 300-319, 2020.

LESK, Michael. **The seven ages of information retrieval**. Ottawa: IFLA, 1996.18p, v.5. Disponível em : https://archive.ifla.org/VI/5/op/udtop5/udt-op5.pdf. Acesso em 5 abr. 2020.

LIU, Ying-Hsang; WACHOLDER, Nina. Evaluating the impact of MeSH (Medical Subject Headings) terms on different types of searchers. **Information Processing & Management**, v. 53, n. 4, p. 851–870, jul. 2017.

LUHN, H. P. A new method of recording and searching information. **American Documentation**, v. 4, n. 1, p. 14–16, jan. 1953.

LUHN, H. P. A Statistical Approach to Mechanized Encoding and Searching of Literary Information. **IBM Journal of Research and Development**, v. 1, n. 4, p. 309–317, out. 1957.

MACULAN, Benildes Coura Moreira dos Santos; AGANETTE, Elisângela Cristina. Desambiguação de relações em tesauros e o seu reúso em ontologias. **Ciência da Informação,** Brasília, v. n.1, n. v.46, p. 18, 2017.

MACULAN, Benildes Coura Moreira dos Santos. **Referências**. [mensagem pessoal]. Mensagem recebida por < erfelipee@gmail.com> em 09 set. 2020.

MANNING, Christopher D.; RAGHAVAN, Prabhakar; SCHÜTZE, Hinrich. **Introduction to information retrieval**. New York: Cambridge University Press, 2008.

MAZZOCCHI, Fulvio. Knowledge organization system (IEKO). **Knowledge Organization**, v. 45, n.1, p.54-78. jan. 2019. Disponível em: https://www.isko.org/cyclo/kos. Acesso em: 6 abr. 2020.

MENDES, Paula Raphisa; REIS, Raquel Martins dos; MACULAN, Benildes Coura Moreira dos Santos. Tesauros no acesso à informação: uma retrospecção. **Revista ACB:** Biblioteconomia em Santa Catarina, Florianópolis, v. 20, n. 1, p. 18, 2015.

MENDONÇA, Fabrício Martins; ALMEIDA, Maurício Barcellos. OntoForInfoScience: A detailed methodology for construction of ontologies and its application in the blood domain. **Brazilian Journal of Information Science**, Marília, v. 10, n. 1, 1 mar. 2016. Disponível em: http://revistas.marilia.unesp.br/index.php/bjis/article/view/5426.Acesso em 09 out 2020.

MÜLLER, Hans-Michael; KENNY, Eimear E; STERNBERG, Paul W. Textpresso: An Ontology-Based Information Retrieval and Extraction System for Biological Literature. **PLoS Biology**, v. 2, n. 11, p. e309, 21 set. 2004.

MOOERS, Calvin N. The next twenty years in information retrieval; some goals and predictions. **American Documentation**, v. 11, n. 3, p. 229–236, ago. 1960.

MOOERS, Calvin N. Zatocoding and Developments in Information Retrieval. **Aslib Proceedings**, v. 8, n. 1, p. 3–22, jan. 1956.

MOOERS, Calvin N. **The Theory of Digital Handling of Non-numerical Information And Its Implications to Machine Economics**. Boston: Zator Co., 1950. Disponível em: https://catalog.hathitrust.org/Record/001691818. Acesso em: 22 maio 2019.

MOTIK, Boris. (Ed.). **OWL 2 Web Ontology Language Profiles**: W3C Recommendation 11 December 2012. 2nd edition. [S.L.]: W3C, 2009.43p. Disponível em: https://www.w3.org/TR/owl2-profiles/. Acesso em: 30 maio 2018.

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

NELSON, Stuart J.; JOHNSTON, W. Douglas; HUMPHREYS, Betsy L. Relationships in Medical Subject Headings (MeSH). *In*: BEAN, Carol A.; GREEN, Rebecca (Orgs.). **Relationships in the Organization of Knowledge**. Dordrecht: Springer Netherlands, 2001. p. 171–184. v. 2. (Information Science and Knowledge Management). Disponível em: http://link.springer.com/10.1007/978-94-015-9696-1_11. Acesso em: 19 jun. 2020.

PAO, Miranda Lee; GREFSHEIM, Suzanne F.; BARCLAY, Mel L.; WOOLLISCROFT, James O.; MCQUILLAN, Mark; *et al*. Factors Affecting Students' Use of MEDLINE. **Computers and Biomedical Research**, v. 26, n. 6, p. 541–555, 1 dez. 1993.

PRIEM, Jason. Beyond the paper. **Nature**, v. 495, p. 4, mar. 2013.

POWERS, Shelley. **Practical RDF**. Beijing; Sebastopol: O'Reilly, 2003.

QIU, Yonggang; FREI, H P. Concept Based Query Expansion. In: SIGIR '93: **Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval**, 1993, p.160-169. Disponível em: https://dl.acm.org/doi/10.1145/160688.160713. Acesso em: 10 mar. 2020.

RECTOR, Alan; SOTTARA, Davide. Formal Representations and Semantic Web Technologies.In: GREENES, Robert. (Ed.). **Clinical Decision Support:** The Road to Broad Adoption.2nd ed. 2. Amsterdã: Elsevier, 2014. p. 551–598. Disponível em: http://linkinghub.elsevier.com/retrieve/pii/B9780123984760000208. Acesso em: 15 abr. 2018.

ROBERTS, Norman. The pre-history of the information retrieval thesaurus. **Journal of Documentation**, v. 40, n. 4, p. 271–285, abr. 1984.

ROBERTSON, S. E.; JONES, K. Sparck (May 1976). "Relevance weighting of search terms". **Journal of the American Society for Information Science**. v.27, n.3, p.129–146.

ROMANO, Lisa. Using Medical Subject Headings (MeSH) in Cataloging. **Technical Services Quarterly**, v. 35, n. 2, p. 217–219, 3 abr. 2018.

RUSSELL, Stuart J.; NORVIG, Peter; DAVIS, Ernest. **Artificial intelligence:** a modern approach. 3rd ed. Upper Saddle River: Prentice Hall, 2010. (Prentice Hall series in artificial intelligence).

SALTON, Gerald. A new comparison between conventional indexing (MEDLARS) and automatic text processing (SMART). **Journal of the American Society for Information Science**, v. 23, n. 2, p. 75–84, mar. 1972.

SALTON, Gerald. A new horizon for information science. **Journal of the American Society for Information Science**, v.47, n.4, p. 333-333, 1996.

SALTON, Gerald. **Automatic information organization and retrieval**. New York: McGraw-Hill computer science series, 1968.

SALTON, Gerard; MCGILL, Michael J. **Introduction to modern information retrieval**. New York: McGraw-Hill computer science series, 1983.

SÁNCHEZ, Miriam Fernández; AZPILICUETA, Pablo Castells. Semantically enhanced Information Retrieval: an ontology-based approach. **Journal of Web Semantics**, v.9, n.4, p.434-452, dec 2011.

SANDERSON, M.; CROFT, W. B. The History of Information Retrieval Research. **Proceedings of the IEEE**, v. 100, n. Special Centennial Issue, p. 1444–1451, maio 2012.

SARACEVIC, Tefko. Ciência da informação: origem, evolução e relações. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 1, n. 1, p. 41-62, jan./jun. 1996.

SARACEVIC, Tefko. An essay on the past and future (?) of information science education—I. **Information Processing & Management**, v. 15, n. 1, p. 1–15, jan. 1979. https://doi.org/10.1016/0306-4573(79)90002-5.

SARACEVIC, Tefko; KANTOR, Paul; CHAMIS, Alice Y; TRIVISON, Donna. A study of information seeking and retrieving. I. Background and methodology. **The Journal of the Association for Information Science and Technology**, v. 39, n. 3, p. 161-176, 1988.

SHIRI, Ali; REVIE, Crawford. Query expansion behavior within a thesaurus-enhanced search environment: A user-centered evaluation. **Journal of the American Society for Information Science and Technology**, v. 57, n. 4, p. 462–478, 15 fev. 2006.

SCHÜTZE, Hinrich; PEDERSEN, Jan O. A co-occurrence-based thesaurus and two applications to information retrieval. **Information Processing & Management**, v. 33, n. 3, p. 307–318, maio 1997.

SHULTZ, Mary. Mapping of medical acronyms and initialisms to Medical Subject Headings (MeSH) across selected systems. **Journal of the Medical Library Association**, v.94, v.4, p. 410–414, out. 2006.

SILBERSCHATZ, Abraham; KORTH, Henry F.; SUDARSHAN, S. **Database system concepts**. Seventh ed. New York, NY: McGraw-Hill, 2020.

SILVA, Daniela Lucas da; SOUZA, Renato Rocha; ALMEIDA, Maurício Barcellos. Ontologias e vocabulários controlados: comparação de metodologias para construção. **Ciência da Informação**, Brasília, v. 37, n. 3, p. 60–75, dez. 2008.

SILVA, Daniela Lucas da; SOUZA, Renato Rocha; ALMEIDA, Maurício Barcellos. Ontologies and Controlled Vocabulary: Comparison of Building Methodologies. In: SMOLNIK, Stefan ; TEUTEBERG, Frank ;THOMAS, Oliver. (Eds). **Semantic Technologies for Business and Information Systems Engineering: Concepts and Applications**. Hershey, PA: IGI Global, 2012.cap.1, p.1-15. Disponível em: http://services.igi-

global.com/resolvedoi/resolve.aspx?doi=10.4018/978-1-60960-126-3. Acesso em 09 out 2020.

SOUZA, Renato Rocha. Sistemas de recuperação de informações e mecanismos de busca na web: panorama atual e tendências. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 11, n. 2, p. 161–173, ago. 2006.

SOUZA, Renato Rocha, TUDHOPE, Douglas; ALMEIDA, Mauricio B. Towards a Taxonomy of KOS: Dimensions for Classifying Knowledge Organization Systems. **Knowledge Organization,** v. 39, n.3, p. 179–192, 2012.

SPARCK JONES, Karen; WILLETT, Peter (Orgs.). **Readings in information retrieval.** San Francisco, Calif: Morgan Kaufman, 1997 (Morgan Kaufmann series in multimedia information and systems).

STOCKER, Markus; PRINZ, Manuel; ROSTAMI, Fatemeh; KEMPF, Tibor. 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, 2018.

SVENONIUS, Elaine. Unanswered questions in the design of controlled vocabularies. **Journal of the American Society for Information Science**, v.36, n. 5, p. 331-340, 1986.

SWANSON, E. Burton; LARSON, Ray R. **Understanding information retrieval systems: management, types, and standards**. Boca Raton, FL: CRC Press, Taylor & Francis Group, 2012.

TOTH, Renato Molina. **Abordagem NoSQL:** uma real alternativa.2011.6p.Disponível em : https://dcomp.sor.ufscar.br/verdi/topicosCloud/nosql_artigo.pdf. Acesso em 09 out 2020.

TUDHOPE, Douglas; BINDING, Ceri. 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.

TURING, Alan Mathison. Computing Machinery and Intelligence. **Mind**, v. 59, n. 236, p. 433–460, 1950.

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

VICKERY, B. C. Ontologies. **Journal of Information Science**, v. 23, n. 4, p. 277–286, ago. 1997.

VIEIRA, Jessica Monique de Lira; SANTOS, Monick Trajano dos; LAPA, Remi Correia. Estudo da construção e aplicação do tesauro na recuperação da informação de teses e dissertações do programa de pós-graduação em desenvolvimento urbano. In: ENCONTRO NACIONAL DE ESTUDANTES DE BIBLIOTECONOMIA, DOCUMENTAÇÃO, GESTÃO E CIÊNCIA DA INFORMAÇÃO, ENEBD, XXXIII.,2010, João Pessoa,

Biblionline, João Pessoa, n. esp., p.71-80, Disponível em: https://periodicos.ufpb.br/ojs2/index.php/biblio/article/view/9629/5238.Acesso em: 09 out 2020.

VOORHEES, Ellen. Query Expansion Using Lexical-Semantic Relations. In: CROFT B.W., VAN RIJSBERGEN, C.J. (eds). In: **Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval.** SIGIR '94. London:Springer, 1994. p. 61–69. https://doi.org/10.1007/978-1-4471-2099-5_7.

WALKER, David. **Query Expansion using Thesauri**: Previous Approaches and Possible New Directions. Los Angeles: University of California, jun. 2001. Disponível em: http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=AD2B077D400DCF1CDD4AFDA522B52C26?doi=10.1.1.94.9929&rep=rep1&type=pdf.Acesso em: 29 jun. 2020.

XU, Jinxi; CROFT, W. Bruce. Query expansion using local and global document analysis. In: **Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval**. 1996. Zurich, Switzerland: Association for Computing Machinery, 1996. p. 4–11. Disponível em: https://doi.org/10.1145/243199.243202. Acesso em: 29 jun. 2020.

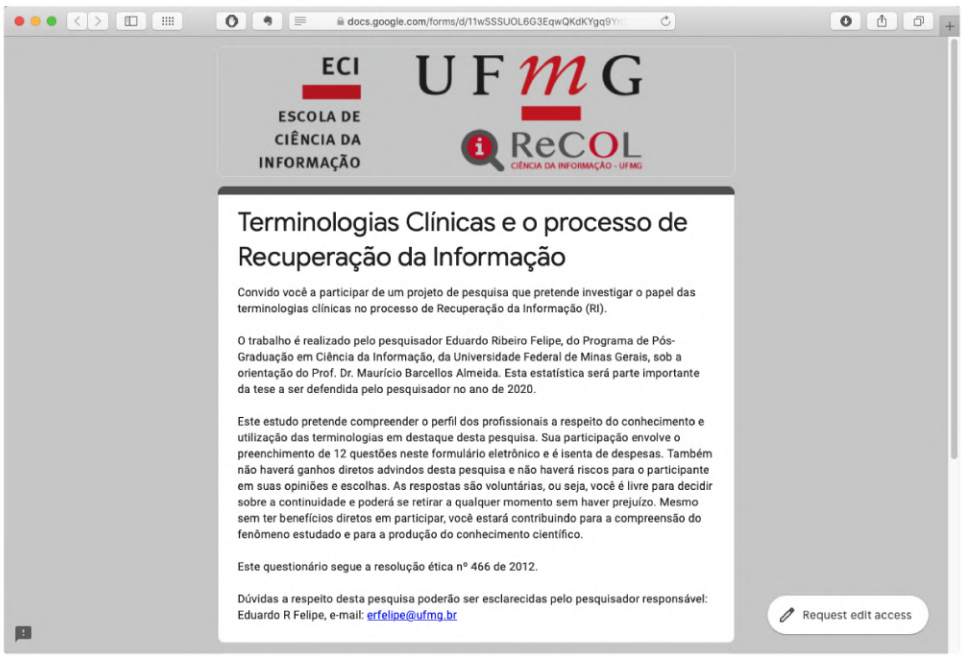
ZIVALJEVIC, Aleksandar; ATALAG, Koray; WARREN, James. Utility of SNOMED CT in automated expansion of clinical terms in discharge summaries: Testing issues of coverage. **Health Information Management Journal**, p. 183335832093452, 21 jul. 2020.

APÊNDICES

APÊNDICE A - Formulário perfil profissional

Objetivando conhecer o perfil dos profissionais que lidam com terminologias em seu ambiente profissional, e seu conhecimento a respeito das terminologias escolhidas para a realização deste trabalho, foi criado um formulário eletrônico para receber as respostas dos participantes, conforme ilustrado pela figura 33.

Figura 31 - Formulário de pesquisa sobre terminologias



Fonte: Dados da pesquisa.

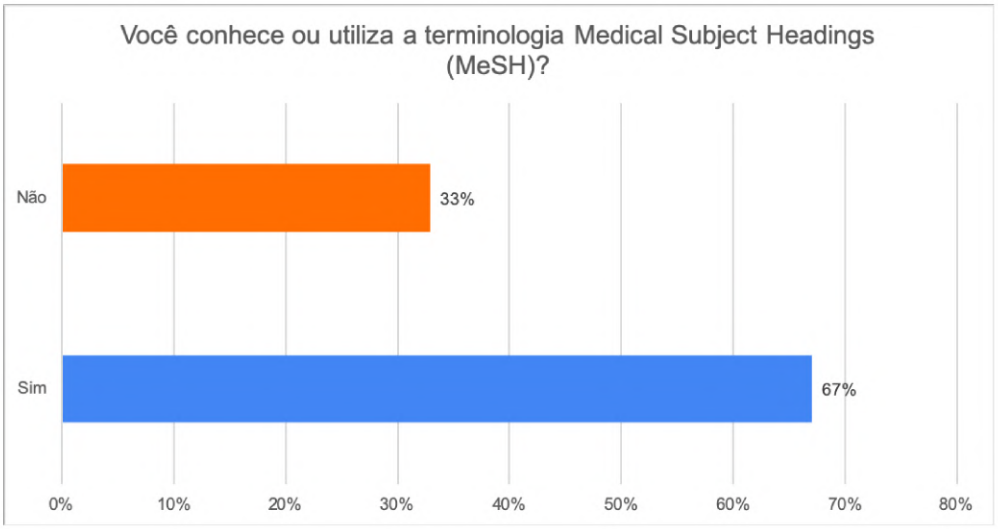
Diversas instituições foram contatadas para a divulgação deste questionário para seus membros e afiliados, entre as quais FEBAB - Federação Brasileira de Associações de Bibliotecários, Cientistas de Informação e Instituições, Sistemas de Bibliotecas da UFMG, grupos de discussão na Espanha e redes sociais de grupos de profissionais deste setor.

Os dados foram importantes para melhor compreensão dos aspectos de utilização prática destas terminologias no ambiente profissional. O formulário[50] de pesquisa sobre a terminologias em CI foi criado para identificar diversas informações pertinentes às terminologias usadas no projeto. Foi projetado em formato eletrônico, na plataforma Google

[50] https://forms.gle/5pKTc3b6tj1uKvkV7

Forms[51]. Era composto de 12 questões, sendo 10 questões fechadas, e 2 abertas. Esteve disponível no período de maio, junho e julho de 2020, sendo divulgado a diversas instituições ligadas a profissionais da Ciência da Informação, bem como em grupos de redes sociais. Este formulário reportou dados dignos de nota em questões que envolviam o conhecimento sobre as terminologias clínicas na comunidade da ECI. A população que participou do questionário, notoriamente, parece conhecer a MeSH e desconhecer, em proporção similar, a SNOMED CT (Figuras 30 e 31).

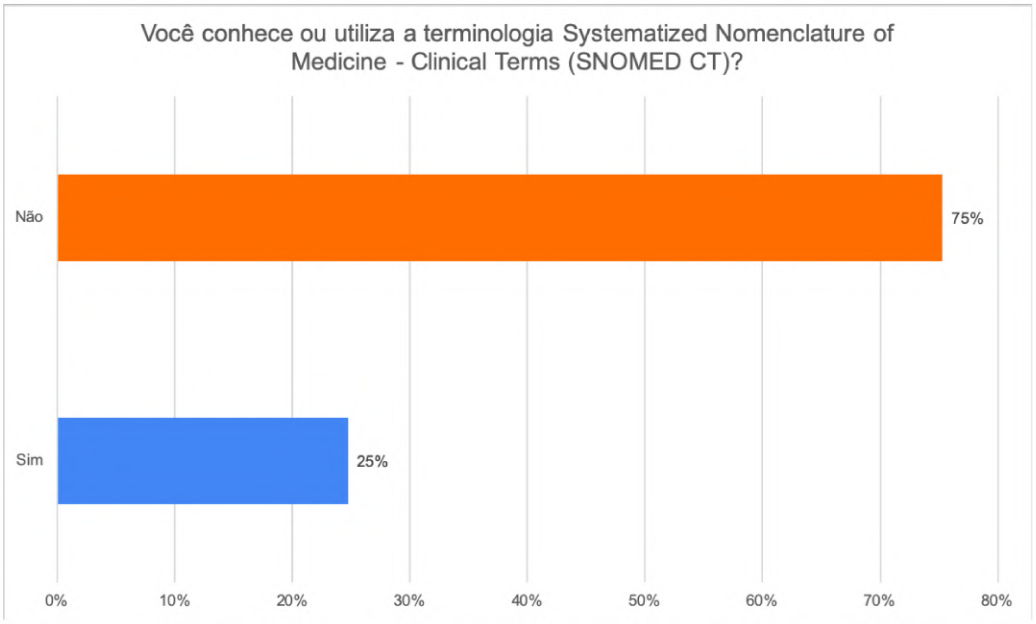
Figura 32 - Conhecimento da terminologia MeSH



Fonte: Dados da pesquisa.

[51] https://www.google.com/forms/about/

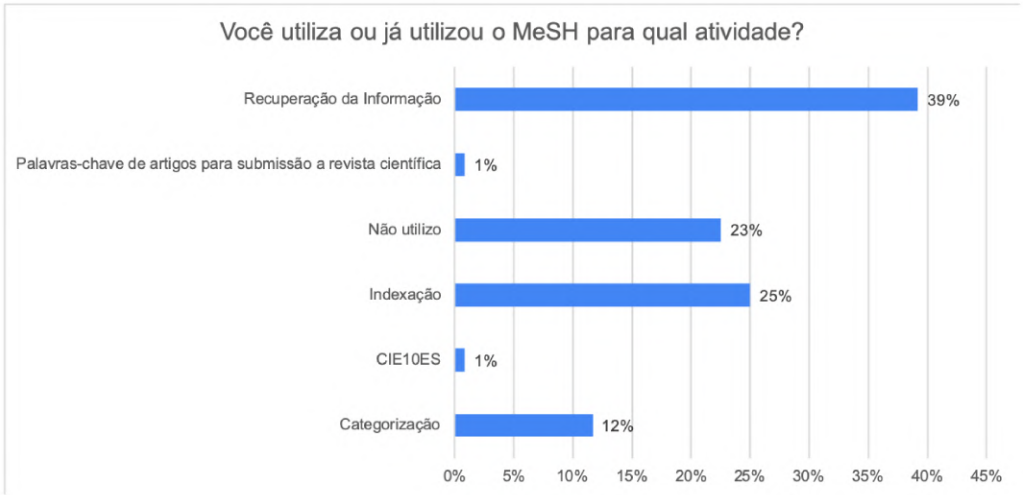
Figura 33 - Conhecimento da terminologia SNOMED CT



Fonte: Dados da pesquisa.

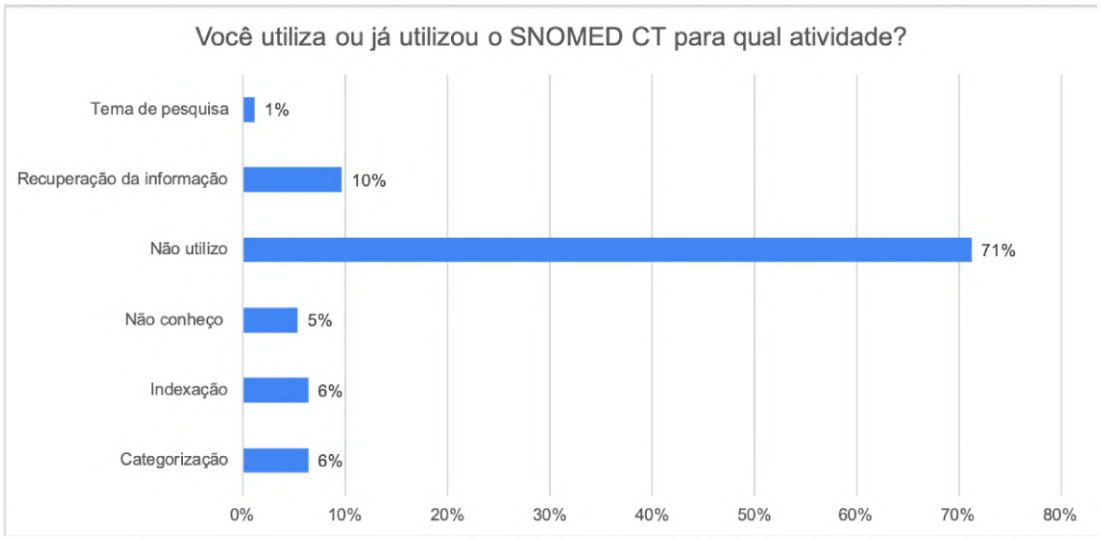
No questionamento sobre a utilização da terminologia, bem como sua finalidade, a pesquisa apontou, a partir das respostas, respectivamente: i) MeSH: predominante a terminologia é usada em RI, com quase 40% das escolhas, seguido pelas atividades de indexação e categorização; cabe notar que, mesmo conhecendo a ferramenta, pelo questionamento anterior, 23% das respostas retratou a não utilização da MeSH; ii) SNOMED CT: o mesmo perfil anterior é percebido, quando 71% das respostas foram marcadas como "Não utilizam" a terminologia, seguido por 10% voltados à RI, com uma pequena participação de 6% para indexação e categorização.

Figura 34- Utilização da MeSH



Fonte: Dados da pesquisa.

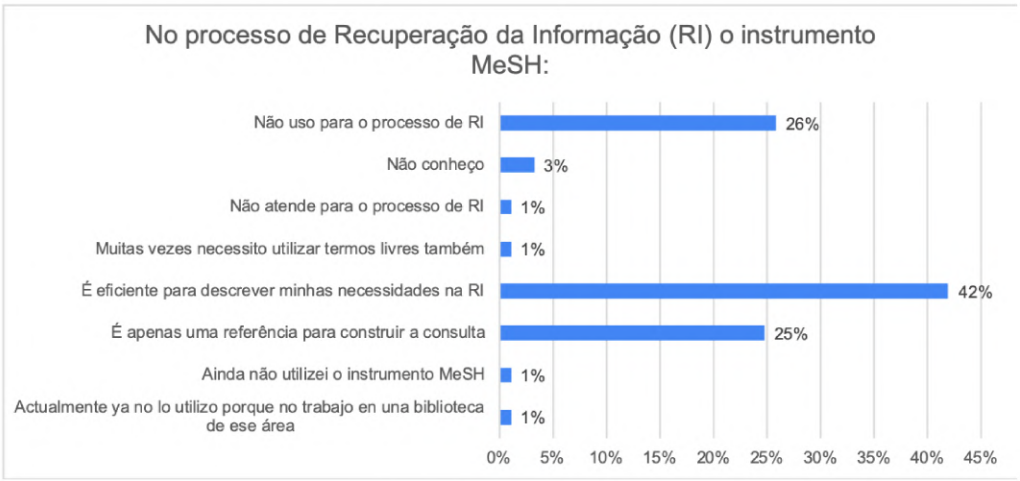
Figura 35 - Utilização da SNOMED CT



Fonte: Dados da pesquisa.

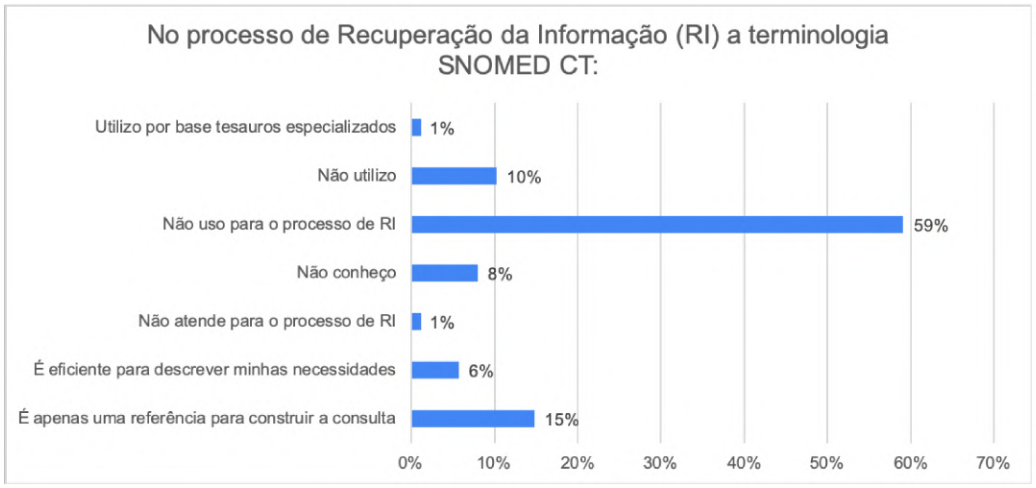
Especificamente sobre o processo de RI, as respostas revelaram que: i) MeSH: a terminologia "é eficiente para descrever minhas necessidades de RI" com 42% das respostas, seguido pela não utilização para esse fim (26%) e "a terminologia é apenas uma referência" para a construção da consulta (25%); ii) SNOMED CT: 59% selecionaram que "não usam a terminologia para o processo de RI", seguido da opção "é apenas uma referência para construir a consulta" (15%), e "não utilizo" e "não conheço" com 10% e 8% respectivamente.

Figura 36 - MeSH e RI



Fonte: Dados da pesquisa.

Figura 37 - SNOMED CT e RI



Fonte: Dados da pesquisa.

Outro questionamento dizia respeito ao conhecimento ou a utilização de outras terminologias. Houve um nítido predomínio do vocabulário DeCS (40%), o que traz à reflexão um ponto importante. Não por acaso a terminologia MeSH é mais conhecida, uma vez que o vocabulário DeCS é uma extensão dessa terminologia. Trata-se de uma iniciativa que permite regionalizar e adicionar termos e relações que não estão presentes na MeSH. Possui tradução também para espanhol, característica que foi expressa em uma resposta de um usuário daquele país. Essas respostas levaram-nos a concluir que a MeSH possui uma aceitação maior pelas seguintes características:

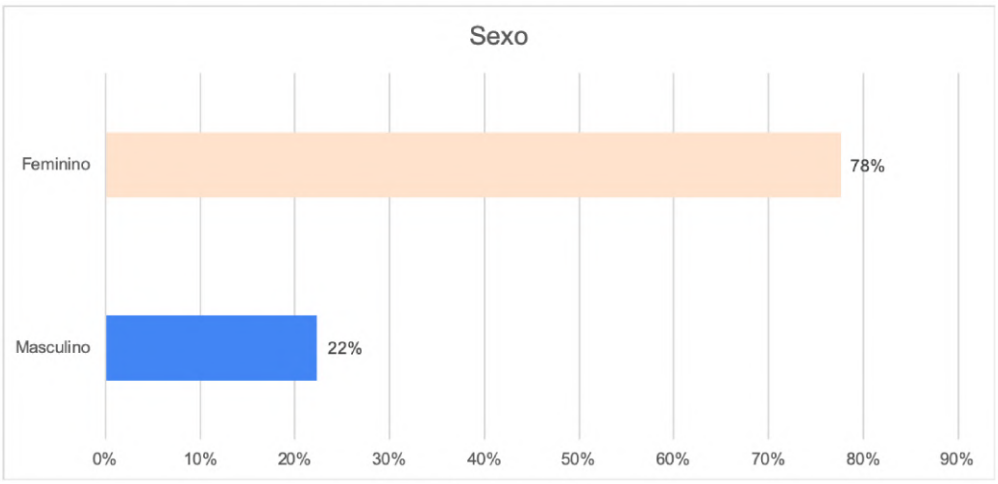
- Possui tradução para diversos idiomas, incluindo Português;
- Recebeu projetos de extensão como o DeCS, nos idiomas português e espanhol;
- Foi uma terminologia concebida com foco em indexação e categorização.

Figura 38 - Outras terminologias conhecidas pelos participantes

Fonte: Dados da pesquisa.

As próximas questões procuraram expressar o perfil dos participantes no questionário, e ajudam a compreender as áreas de atuação, perfil pessoal e profissional. A maioria das pessoas que contribuiu voluntariamente no questionário são mulheres.

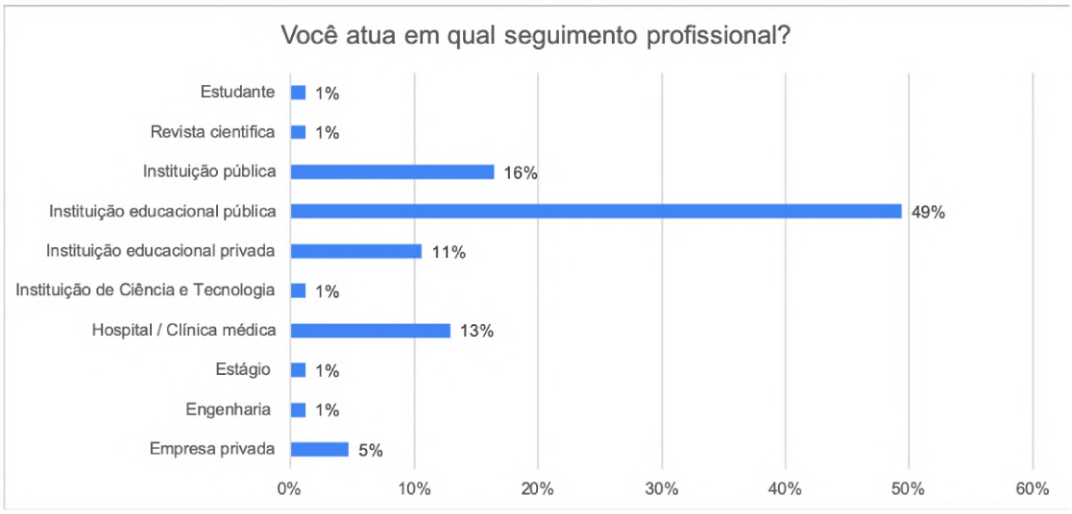
Figura 39 - Sexo dos participantes



Fonte: Dados da pesquisa.

Com relação à área de atuação profissional, praticamente a metade, 49%, trabalham em “instituição educacional pública”, seguido por “instituição pública” (16%), “hospital ou clínica médica” (13%) e “instituição educacional privada” (11%).

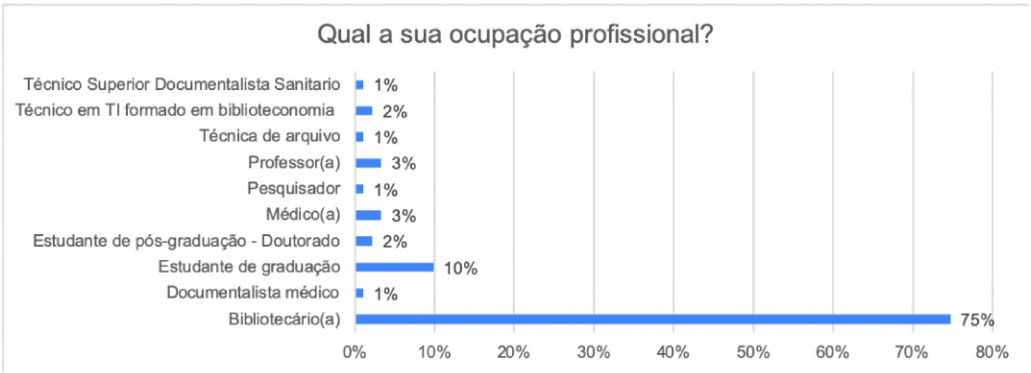
Figura 40 - Seguimento profissional dos participantes



Fonte: Dados da pesquisa.

Em relação a ocupação profissional, as respostas demonstraram maciça participação de bibliotecários (75%), estudantes de graduação (10%), professores (3%) e médicos (3%).

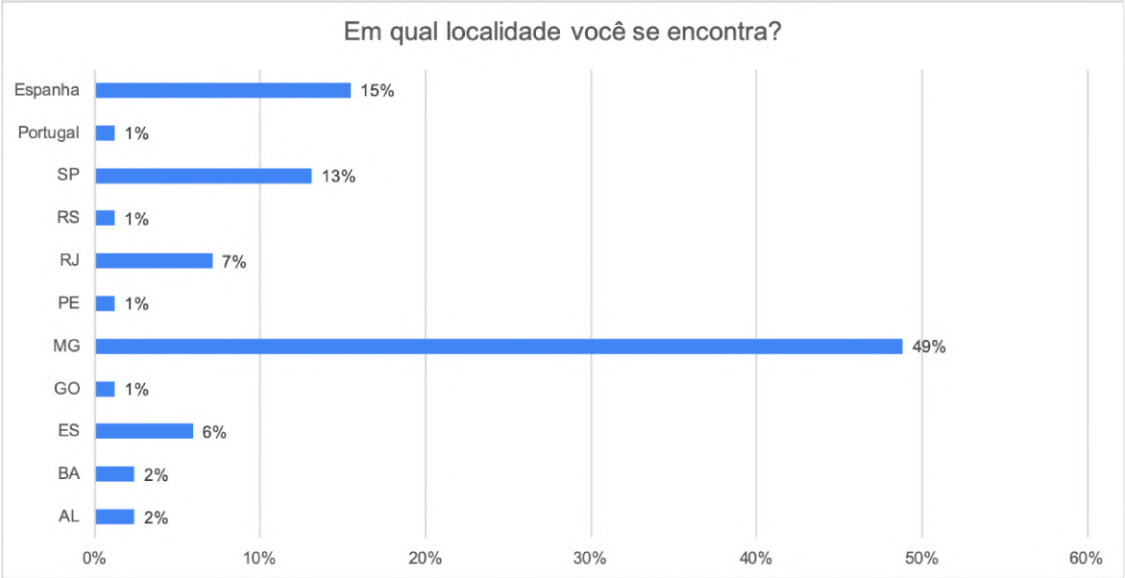
Figura 41 - Ocupação profissional dos participantes



Fonte: Dados da pesquisa.

A localização geográfica dos participantes foi diagnosticada da seguinte forma: 84% dos participantes estão no Brasil, predominantemente em Minas Gerais (49%), seguido por São Paulo (13%), Rio de Janeiro (7%) e Espírito Santo (6%) como participações de maior expressividade. A pesquisa ainda contou com participações da Espanha (15%) e Portugal (1%), sendo esses últimos alinhados ao perfil brasileiro no que diz respeito a maior utilização da MeSH em contraste com a SNOMED CT.

Figura 42 - Localidade geográfica



Fonte: Dados da pesquisa.

O questionário foi um importante instrumento para a percepção mais direta daquilo que o andamento da pesquisa indicava por suposição a respeito da maior aceitação da terminologia MeSH.